Theor Appl Genet (2013) 126:1419–1430
DOI 10.1007/s00122-013-2100-2

REVIEW

Family-based association mapping in crop species

Baohong Guo · Daolong Wang · Zhigang Guo · William D. Beavis

Received: 8 November 2012 / Accepted: 2 April 2013 / Published online: 26 April 2013

Abstract Identification of allelic variants associated with complex traits provides molecular genetic information associated with variability upon which both artificial and natural selections are based. Family-based association mapping (FBAM) takes advantage of linkage disequilibrium among segregating progeny within crosses and among parents to provide greater power than association mapping and greater resolution than linkage mapping. Herein, we discuss the potential adaption of human family-based association tests and quantitative transmission disequilibrium tests for use in crop species. The rapid technological advancement of next generation sequencing will enable sequencing of all parents in a planned crossing design, with subsequent imputation of genotypes for all segregating progeny. These technical advancements are easily adapted to mating designs routinely used by plant breeders. Thus, FBAM has the potential to be widely adopted for discovering alleles, common and rare, underlying complex traits in crop species.

Communicated by R. Varshney.

B. Guo (✉)
Syngenta Biotechnology, Inc, 2369 330th Street, Slater, IA 50244, USA
e-mail: baohong.guo@syngenta.com

D. Wang · Z. Guo
Syngenta Biotechnology, Inc, 3054 Cornwall Road, Research Triangle Park, NC 27709, USA

W. D. Beavis (✉)
Iowa State University, 1208 Agronomy Hall, Ames, IA 50011, USA
e-mail: wdbeavis@iastate.edu

Introduction

The purposes of identifying allelic variants associated with complex traits are to understand the molecular mechanisms of complex traits and provide diagnostic and selectable markers for favorable alleles in genetic improvement. Three forward genetics approaches have been proposed to serve this purpose: linkage mapping (also called QTL mapping), association mapping (AM) and family-based association mapping (FBAM). Linkage mapping uses a single family (Lander and Bostein 1989) or several families (Xu 1998; Blanc et al. 2006) of segregating progeny from cross(es) between inbred lines, but is constrained by a limited number of recombination events created in production of segregating progeny, resulting in high power but low resolution (Darvasi et al. 1993). Association mapping uses unrelated individuals which have accumulated historical recombination events for a large number of generations, thus improving the resolution of causal variants (Anderson and Georges 2004; Yu et al. 2006). However, large samples (thousands or more) are required to provide sufficient power to identify less frequent alleles with significant impact on the trait of interest (Hirschhorn and Daly 2005; Kingsmore et al. 2008).

The use of several families in linkage mapping is akin to FBAM in that multiple families of segregating progeny are obtained from matings among lines. However, there are distinctions. In FBAM it is assumed that parental lines are related through identity by descent (IBD) and linkage disequilibrium (LD) of alleles (historic recombination events) in parental lines is exploited to provide greater resolution for identifying the variants associated with complex traits (Fig. 1), whereas in multiple family linkage mapping, parental lines in independent families (Xu 1998) or connected families (Blanc et al. 2006) are assumed to be

unrelated and alternative alleles are modeled as members of distinct haplotypes in these parental lines, resulting in an assumption of no LD of alleles among parental lines and therefore a lower resolution of QTL detection. Often multiple family linkage mapping is conducted with sparse genotyping technologies (Li et al. 2005; Guo et al. 2006; Buckler et al. 2009) whereas FBAM requires technologies capable of providing greater densities of marker loci (Yu et al. 2008; Guo et al. 2010; Tian et al. 2011; Kump et al. 2011). There is no clear definition to distinguish sparse and dense genotyping, but successful identification of QTLs in FBAM is possible only when genotyped loci are in linkage disequilibrium (LD) with causal variants. And LD decay varies among crops and populations within crops (Buckler and Gore 2007). In human genetic research, FBAM is less attractive than AM due to costly recruitment of consenting family members (Laird and Lange 2006) and it may have a limited power for detecting QTLs because of small numbers of progeny per family. Plant and experimental animal species (Churchill et al. 2004), however, exhibit qualities that favor application of FBAM. For example, parental lines could be fully inbred lines and large segregating families are relatively easy to develop with designed matings. Because plant breeders of most crop species typically mate a few elite inbred lines or varieties with a wide range of new inbred lines or varieties to generate a large number of segregating populations (Jansen et al. 2003), FBAM may be applied to established breeding populations in breeding programs.

Two approaches are used to exploit LD of alleles among parental lines in FBAM. One is to test marker loci which are genotyped in a FBAM population or genotyped in parental lines and imputed onto members of a FABM population (see below). In this approach, markers adjacent to or within a QTL gene may show a genetic effect of a QTL since it resides on or is in complete LD with the causal variant of the QTL or a reduced genetic effect due to its incomplete LD with the causal variant of the QTL in parental lines. Association of markers with a trait may be significantly detected even if it is in incomplete LD with causal variant of a QTL in parental lines as long as an appropriately large FBAM population size is used. The other one is to test a series of unobservable loci along a chromosome or a target region in which IBD probabilities of a QTL allele among parental lines at the locus being tested are first estimated based on the marker haplotype similarities of parental lines and subsequently combined with IBD probabilities of a QTL allele estimated using its flanking markers within known pedigrees (George et al. 2000; Meuwissen et al. 2002; Farnir et al. 2002; Lund et al. 2003; Lee and Werf 2004). The former approach generally requires a higher density of genotyping than the latter, and both approaches need a higher marker density than multiple family linkage mapping. In this review, we only focus on the first approach since emergence of next generation sequencing technologies enables genotyping by sequencing (GBS) and identification of most potential allelic variants for parental lines.

Nested association mapping (NAM) populations, in which a set of diverse lines is crossed with a common reference line, have been set up for several plant species including maize (Yu et al. 2008), Arabidopsis (Buckler and Gore 2007), barley (R. Wise, personal communication), sorghum (J. Yu, personal communication), and soybean (B. Diers, personal communication). These NAM populations have been referred to as next generation mapping populations (Morrell et al. 2012). We propose that NAM represents a logical step toward FBAM which will provide a cost-effective approach to discover a wide spectrum of alleles, common or rare, underlying complex traits in plant species.

Creation of a FBAM population

Mating designs

In contrast to human species, a FBAM population consisting of multiple families may be developed in plant species by designed matings involving fully inbred lines. The power of QTL detection will depend upon the total number of segregating progeny sampled from informative families (Allison et al. 1999). Generally, variants which have high minor allele frequencies (close to 0.5) will segregate in about half of the families, whereas variants of low minor allele frequencies will be found in either very few or a large number of informative families (Guo et al. 2010). Therefore, an important goal of FBAM mating designs is to assure that variants of low minor allele frequencies are included in a reasonable number of informative families. One well-known design for producing a FBAM population is single reference mating design, i.e., the NAM design (Yu et al. 2008). In this design, a diverse set of parental lines is selected based on molecular diversity analysis, and each of the lines is crossed with a common reference line. A second design is the multiple references mating design (NCD-1) proposed by Guo et al. (2010). In this design, a set of diverse parental lines is divided into several groups and then crossed with one of several reference lines. Compared with the NAM design, NCD-1 enables more variants to be evaluated in a reasonable number of informative families. A third choice could be to develop families from matings between lines of contrasting phenotypes. This design is based on the assumption that alleles underlying the specific trait may be enriched for alternative alleles in the extreme phenotypes.

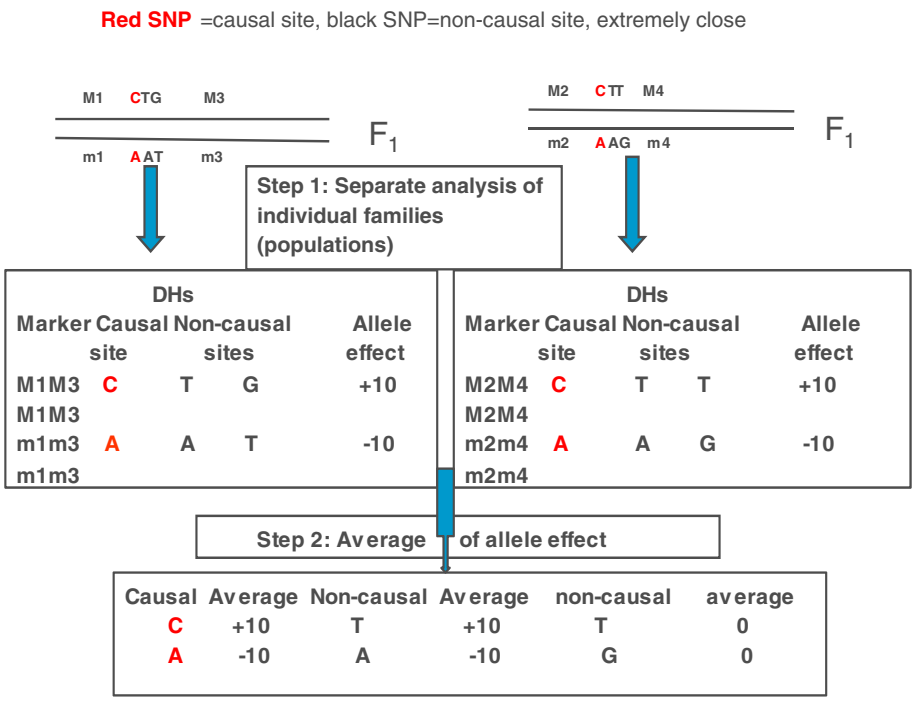

Fig. 1 Family-based association mapping (FBAM) exploits linkage disequilibrium in parental lines to identify the variants associated with a complex trait. For illustration, assume two families consisting of double haplotype (DH) progeny from crosses between fully inbred lines. Also assume that three single nucleotide polymorphism (SNP) loci (C/A, T/A, T/G) are in close proximity on one chromosome. Further assume that the C/A locus (*red highlighted*) is a functional locus underlying the complex trait, with genetic effects of C and T alleles being +10 and −10 separately. T/A and G/T loci are assumed to be non-causal SNPs. The former one is in complete linkage disequilibrium (LD) but the latter one in linkage equilibrium (LE) with the causal SNP. These three loci are assumed to be flanked by two close polymorphic markers M1/m1 and M3/m3 in the first family and by M2/m2 and M4/m4 in the second family. Recombination events among the SNP loci and double-cross over events between two flanking markers are not expected to occur due to the above assumptions when the DH progeny are produced. Individuals with recombinants between two flanking markers are not considered here for convenient illustration. Conceptually, FBAM can be decomposed into two steps: (1) analysis of individual families to estimate the genetic effects of marker loci, and (2) average of estimated genetic effects across families for each marker locus. For illustration, consider environmental and experimental errors are zero. The T/A locus, which is in complete LD with the causal C/A, has the same average genetic effect of +10/−10 as the causal C/A and therefore both loci are indistinguishable. The T/G locus, which is in equilibrium with the causal C/A, has an average genetic effect of 0 and it is distinguished from the causal C/A. Partial linkage disequilibrium and other sources of variability will add complexity and a one-step sophisticated statistical analysis is required to detect the variants underlying complex traits in FBAM. In multiple family linkage mapping, genotypes of the three SNP loci are unknown and these three loci are not distinguishable, resulting in a reduced resolution

This design combined with the extreme trait sequencing may maximize power for detecting the less frequent alleles underlying complex traits.

Types of progeny

Theoretically, any type of segregating progeny can be used for identifying QTLs using FBAM (Guo et al. 2010). However, RILs and DHs are genetically reproducible, easily stored and can be used for precise phenotyping multiple traits at multiple locations and years.

FBAM not only captures historic recombination events in parental lines but also produces a new set of recombinants in the FBAM population. Such recombinants help to eliminate spurious long range LD on a chromosome (Guo et al. 2010) while LD within short segments of chromosomes formed by historic recombination events will be maintained. Because RILs produced by selfing accumulate twice as many recombinant events as DH and F_2, RILs are more desirable to reduce LD over large segments of chromosomes and improve resolutions of QTLs.

Sample sizes of parental lines and progeny

Segregating QTLs in a FBAM population are determined by parental lines and therefore an appropriate number of diverse parental lines should be included so that all or most of the variants underlying a complex trait could be included in a FBAM population. The power of detection of a QTL depends upon the number of informative families and the number of segregating progeny sampled from them. Guo et al. (2010) indicated that more than 5 informative

families with 100 lines per family (a total of more than 500 informative individuals) will produce reasonable power for detection of a QTL with genetic effects responsible for at least 5 % of the phenotypic variability. In practice, however, the causal variants underlying a complex trait is unknown prior to development of an experimental FBAM population. One feasible strategy is to enable as many SNPs as possible to have an appropriate number of informative individuals through balancing family sizes and number of families if the SNP genotypes of parental lines are available.

Genotyping and imputing of genotypes in a FBAM population

In early human FBAM studies, a FBAM population and their parental lines are assumed to be genotyped using the same set of high density molecular markers. Recently, Burdick et al. (2006) developed in silico genotyping method for imputing high density genotypes of progeny based on densely genotyped parental lines and sparsely genotyped progeny. Yu et al. (2008) developed a similar strategy for the maize NAM population.

The method developed by Burdick et al. (2006) is applicable for plant species but construction of haplotype phases is not necessary when inbred lines are used as parents. Genotypes of parental lines are directly copied onto their corresponding offspring using flanking markers for their defined genomic segments where no recombination events occur. Missing data are imputed for the interval where a recombination event occurs. The method used by Yu et al. (2008) is similar when no recombination events occur in the segment defined by flanking markers, but for the segment where a recombination event occurs, a recombination event is simulated and then the corresponding parental genotypes are imputed.

Guo and Beavis (2011) developed an expectation imputation method using Haley and Knott's concept (Haley and Knott 1992). First, linkage map positions of unmapped SNPs were interpolated using genomic physical map and known linkage map and then expected values are computed and imputed using a series of linked markers which are genotyped in parental lines and their progeny. This method is applicable for any set of progeny from self-pollinations of a mating between two inbred lines. It has been successfully applied to impute 0.5 million SNP genotypes of 26 parental lines onto the maize NAM population of 5,000 RILs using 1,000 markers (http://www.agron.iastate.edu/GFSPopGen/resources.html). Approximately 70 % of data points had absolute genetic scores of 0.9–1.0 which are close to true genetic score with absolute value of 1. Five percent of data points had expected scores with absolute values below 0.9 due to missing data of 1,000 markers in RIL progeny. About 25 % of data points were missing due to failure of genotyping in parental lines. Recently, DH_{F1} (F_1 derived double haplotype) and DH_{F2} (F_2 derived double haplotype) have become common in maize (Bernardo 2009). This technique can be naturally applied to generate FBAM and NAM populations. Imputation of genotypes can be similarly obtained by use of conditional distributions described by Snape (1988) for DH_{F1} and in Appendix for DH_{F2}, respectively.

Tian et al. (2011) developed a method which computes and imputes the weighted values of parental lines using physical distance instead of genetic distance of markers to be imputed relative to flanking markers. One obvious drawback of this method is that the same genotype scores are imputed for DH_{F1} and RILs although RILs have almost twice as many recombination events as DH_{F1} for a given genomic segment.

Data analysis methods

Family strata of a FBAM population

Family strata cause differences of phenotypic means of families which may cause false positive associations and unbiased estimates of genetic effects in FBAM (Abecasis et al. 2000). In analysis of data from human families, family strata are defined as diverged subpopulations from which different families are drawn. In crop species, family strata may be produced by matings between heterogeneous pairs of inbred lines. Family strata may exist if crosses are made separately within diverged groups of inbred lines or if a NAM mating design is used to produce a FBAM population because the set of parental lines crossed with a common reference line is usually from diverse origins. Similarly, family strata may exist if a FBAM population is produced using a NCD-1 mating design. Strata may not exist if each of the matings is made between inbred lines from the same homogenous group of inbred lines or between two distinct homogenous groups of inbred lines. In addition, strata-like differences may appear among families if the progeny from the same families are evaluated in the same environments while progeny from different families are evaluated in distinct environments.

Two methods are used to control family strata in data analysis. One is to include family mean effects as fixed (Yu et al. 2008; Guo et al. 2010) or random (Abecasis et al. 2000) in the models. The second method is to decompose the genetic score of an offspring into the expected genetic score given parental lines and a deviation from the expected score. The latter is used in human quantitative transmission disequilibrium tests (QTDT) (Abecasis et al.

2000) and family-based association tests (FBAT) (Laird and Lange 2006). Note that both methods are used in QTDT.

Another issue in FBAM data analyses is segregation of background QTLs. There are two methods for controlling background QTLs. One is to use pedigree information (Abecasis et al. 2000). The other one is to use multi-locus model in which QTL-linked molecular markers are included as cofactors in the models (Yu et al. 2008; Valdar et al. 2009; Guo et al. 2010). A combination of both, i.e. inclusion of QTL-linked markers as cofactors to control major effect QTLs and use of pedigree information to control a large number of small effect QTLs, may be the best approach, although further researches are needed to evaluate this idea.

Family-based association tests (FBAT)

Laird and his group (Laird and Lange 2006) developed FBAT for human FBAM. FBAT is non-parametric and its validity does not require specification of the distributions underlying complex traits. The assumption of Mendelian inheritance ensures valid results of test statistics. It is applicable for any trait including discrete and selected traits but is more suitable for a FBAM population with a small number of progeny per family and a large number of families (Laird and Lange 2006).

The general FBAT statistic is defined as:

$$U = \sum_i U_i, \quad \text{where } U_i = \sum_j T_{ij}(X_{ij} - E(X_{ij}|P_i))$$

$$\text{Var}(U) = \sum \text{Var}(U_i), \quad \text{where Var}(U_i) = E(U_i^2) - [E(U_i)]^2$$

$$Z_{\text{FBAT}} = U/\sqrt{\text{Var}(U)} \text{ or } \chi^2_{\text{FBAT}} = U^2/\text{Var}(U)$$

where i indexes a pedigree, j indexes the progeny in the ith pedigree. A pedigree is defined as a nuclear family consisting of progeny from a cross between two parental lines, as described in Table 1, or a set of families in which a progeny line from one family is a parent of another family, as described in Table 2. X_{ij} is the genetic score of progeny line j within family i. In an additive model, X_{ij} is defined as the number of copies of a particular allele. In a recessive model, $X_{ij} = 1$ if the progeny line is A_1A_1 and 0 otherwise. In a dominant model, $X_{ij} = 1$ if the progeny line has any number of A_1 alleles and 0 otherwise. $E(X_{ij} \mid P_i)$ is the expected value of X_{ij} given parental genotypes. $T_{ij} = Y_{ij} - \mu$, where Y_{ij} is the phenotypic value of progeny line j within family i and μ is user-defined. Several choices are available for μ: (a) sample mean, (b) the value that minimizes Var(U), or (c) $T_{ij} = Y_{ij} - \beta_0 - \beta Z_{ij}$, where Z_{ij} are covariates. Given a sufficiently large sample (at least 10 informative families), Z_{FBAT} is distributed as N(0,1) and χ^2_{FBAT} as χ^2_1.

Computation of Var(U) depends upon the distribution of marker genotypes among offspring under the null hypothesis, in which the marker being tested is assumed to be unrelated to the trait and it assumes Mendelian inheritance (Rabinowitz and Laird 2000; Horvath et al. 2004). An algorithm and software have been developed for human samples, but not plant species (http://www.biostat.harvard.edu/fbat). In plant species, the parental lines and their offspring are inbred lines in most cases and parental genotypes are usually available. Tables 1 and 2 illustrate an algorithm to derive the null distributions of marker genotypes among offspring in plant species using a hypothesized nuclear family and a hypothesized pedigree of inbred offspring, respectively. The pedigree described in Table 2 is used to illustrate how to calculate Var(U_i) of a pedigree. The computation of a nuclear family is similar. The observed genotypes of founder parental lines A, B and C and offspring lines O_1 and O_2 in Table 2 are assumed to be separately AA, BB, BB, AA and BB at the locus being tested, with the assumed phenotypic values of offspring lines O_1 and O_2 being 11 and 14, respectively. Further assume that the generations of offspring O_1 and O_2 are F_2, and the sample mean of the whole FBAM population is 10. Use the number of A alleles to define the genetic score (additive model). The genetic scores of O_1 corresponding to the first column of Table 2 is (2, 2, 2, 1, 1, 1, 0) and the genetic scores of O_2 corresponding to the third column of Table 2 (2, 1, 0, 2, 1, 0, 0). The joint O_1 and O_2 null distribution according to the last column of Table 2 are (1/16, 1/8, 1/16, 1/16, 1/8, 5/16, 1/4), with the marginal frequencies of O1 genotypes AA, AB and BB being 1/4, 1/2 and 1/4, respectively (i.e. Prob(O_1 | parental lines A and B)), and the marginal frequencies of O_2 genotypes AA, AB and BB being 1/8, 1/4, 5/8, respectively (i.e. Prob(O_2 | founder parental lines A, B, C). Note that the frequencies of O_1 are the same as the segregating frequencies of a nuclear family because the O_1 family itself is a nuclear family whereas the frequencies of O_2 are not the same as the segregating frequencies of a nuclear family because the genotype of O_2 depends upon the genotype of O_1. The expected genotypic scores of O_1 and O_2 are $E(O_1 \mid A, B) = (1/4)2 + (1/2)1 + (1/4)0 = 1.0$ and $E(O_2|A, B, C) = (1/8)2 + (1/4)1 + (5/8)0 = 0.5$. The observed U_i value of this pedigree is $(11 - 10)(2 - 1.0) + (14 - 10)(0 - 0.5) = -1$. The observed U is the sum of the observed U_i for all the nuclear families and pedigrees segregating at the locus being tested. Note the phenotypic values of O_1 corresponding to each row of Table 2 are the observed O_1 value of 11 and the phenotypic values of O_2

Table 1 A null distribution of offspring genotypes in a hypothesized nuclear family of two offspring lines O_1 and O_2, which are produced by selfing the F_1 progeny from cross between two inbred lines A and B for $(t-1)$ generations (i.e., F_t generation)

O_1 with phenotypic value y_1		O_2 with phenotypic value y_1		Joint O_1 and O_2 null distribution
Genotype	Frequency	Genotype	Frequency	
AA	$0.5–0.5^t$	AA	$0.5–0.5^t$	$(0.5–0.5^t)^2$
		AB	0.5^{t-1}	$(0.5–0.5^t)\ 0.5^{t-1}$
		BB	$0.5–0.5^t$	$(0.5–0.5^t)^2$
AB	0.5^{t-1}	AA	$0.5–0.5^t$	$0.5^{t-1}(0.5–0.5^t)$
		AB	0.5^{t-1}	$0.5^{2(t-1)}$
		BB	$0.5–0.5^t$	$0.5^{t-1}(0.5–0.5^t)$
BB	$0.5–0.5^t$	AA	$0.5–0.5^t$	$(0.5–0.5^t)^2$
		AB	0.5^{t-1}	$(0.5t–0.5^t)0.5^{t-1}$
		BB	$0.5–0.5^t$	$(0.5–0.5^t)^2$

Assume that lines A, B, O_1 and O_2 have observed genotypes AA BB, BB and AA at the locus being tested, respectively. Let the phenotypic values of O_1 and O_2 be y_1 and y_1, respectively. According to Mendel Law, offspring segregates at AA, AB and Bb with frequencies (probabilities) of $0.5–(0.5)^t$, $(0.5)^{t-1}$, $0.5–(0.5)^t$ at tth selfing generation. Prob (1st line = g1, 2nd line = g2) = P(g1) P(g2), where g1, g2 = AA, AB, BB

Table 2 A null distribution of offspring genotypes in a hypothesized pedigree, where one inbred line O_1 is produced by selfing the F_1 progeny from cross between inbred lines A and B for (t_1-1) generations (i.e., F_{t1} generation) and another offspring O_2 from cross between O_1 and inbred line C for (t_2-1) generations (i.e., F_{t2} generation)

O_1 with phenotypic value y_1		O_2 with phenotypic value y_2		Joint O_1 and O_2 null distribution
Genotype	Frequency	Genotype	Frequency	
AA	$0.5–0.5^{t1}$	AA	$0.5–0.5^{t2}$	$(0.5–0.5^{t1})(0.5–0.5^{t2})$
		AB	0.5^{t2-1}	$0.5^{t2}\ (1–0.5^{t1-1})$
		BB	$0.5–0.5^{t2}$	$(0.5–0.5^{t1})(0.5–0.5^{t2})$
AB	0.5^{t1-1}	AA	$0.5\ (0.5–0.5^{t2})$	$0.5^{t1}(0.5–0.5^{t2})$
		AB	0.5^{t2}	$0.5^{t1+t2-1}$
		BB	$0.5 + 0.5\ (0.5–0.5^{t2})$	$0.5^{t1}(1.5–0.5^{t2})$
BB	$0.5–0.5^{t1}$	BB	1	$0.5–0.5^{t1}$

Assume that lines A, B, C, O_1 and O_2 have the observed genotypes AA, BB, BB, AA and BB at the locus being tested, respectively. And Let the phenotypic values of O_1 and O_2 be y_1 and y_2, respectively. The hypothesized pedigree can be broken into two nuclear families. One consists of parental lines A and B and offspring O_1. The other has parental lines O_1 and C and offspring O_2. In the former family, O_1 segregates at AA, AB and BB (Table 1). In the latter family, the null distribution of O_2 depends upon genotype of O_1, which is crossed with parental line C (with genotype BB). Given O_1 with AA, O_2 segregates as in Table 1. Given O_1 with AB, O_2 segregates as would selfing a backcross for $t_2 - 1$ generations. Given O_1 with BB, O_2 does not segregate

the observed O_2 value of 14 according to the FBAT theory. Similarly, the U_i values of a joint O_1 and O_2 null distribution corresponding to each row of Table 2 are $((11-10)(2.0-1.0)+(14-10)(2.0-0.5),\ (11-10)(2.0-1.0)+(14-10)(1.0-0.5),\ \ldots,\ (11-10)(0-1.0)+(14-10)(0-0.5)) = (7, 3, -1, 6, 2, -2, -3)$. The corresponding frequencies of these U_i values are the same as the joint O_1 and O_2 null distribution described above. $E(U_i) = (1/16)7 + (1/8)3 + \cdots + (1/4)(-3) = 0$. $E(U_i^2) = (1/16)7^2 + (1/8)3^2 + \cdots + (1/4)(-3)^2 = 10.5$. $\mathrm{Var}(U_i) = 10.5 - 0^2 = 10.5$. $\mathrm{Var}(U)$ is the sum of $\mathrm{Var}(U_i)$ for all the nuclear families and pedigrees segregating at the locus being tested.

In a method referred to as pedigree disequilibrium tests (PDTs) (Martin et al. 2000; Monks and Kaplan 2000), $\mathrm{Var}(U_i)$ is empirically estimated (Lange et al. 2002). The PDTs have been adapted for plant species and is referred to as quantitative inbred PDT (QIPDT) (Stich et al. 2006).

In both FBAT and PDT, within family information only is used for detecting QTLs. Recently, Steen et al. (2005) developed a two-step procedure in which candidate SNPs are selected using the among family information and they are then tested using FBAT.

Quantitative transmission disequilibrium tests (QTDT)

In contrast to the FBAT, QTDT requires specification of a normal distribution for trait of interest. Abecasis et al. (2000) developed this method based on Fulker et al. (1999).

The most distinct feature of this method is that the genotypic score g_{ij} of an offspring individual j in family i at the marker locus being tested is decomposed into between and within family components b_i and w_{ij}. b_i is the expected genotypic score of an offspring individual given parental genotypes. w_{ij} is a deviation of observed genotypic score from b_i. The relationship between phenotype y_{ij} and marker being tested is described by:

$$y_{ij} = \mu + \beta_b b_i + \beta_w w_{ij}$$

where y_{ij} is the phenotypic value of an offspring individual j in family i. μ is the grand mean. If family strata exist, β_b is a biased estimator whereas β_w provides an unbiased estimate of genetic effects. If family strata do not exist, both are unbiased estimators of genetic effects (Fulker et al. 1999; Abecasis et al. 2000).

The variance–covariance matrix is defined for each family as:

$$\begin{aligned}\Omega_{ijk} &= \sigma_a^2 + \sigma_s^2 + \sigma_e^2 \quad \text{if individuals } j = k \text{ in family } i \\ &= \pi_{ijk}\sigma_a^2 + \sigma_s^2 \quad \text{if individuals } j \neq k \text{ in family } i\end{aligned}$$

Equivalently, the above model can be re-written for crop species as:

$$y_{ij} = \mu + \beta_b b_i + \beta_w w_{ij} + \mu_i + u_{ij} + e_{ij}$$

where y_{ij}, b_i, and w_{ij} as well as μ, β_b, and β_w are described as above. μ_i is family mean effect excluding locus being tested. Let $\mathbf{s}$ be $\{\mu_i\}$, vector of family mean effects. $\mathbf{s} \sim N(\mathbf{0}, \sigma_s^2 \mathbf{I})$, where σ_s^2 is family mean effect variance, $\mathbf{0}$ is zero vector and $\mathbf{I}$ is identity matrix. u_{ij} represents polygenic effects (background genetic effects excluding locus being tested) of an offspring individual i in family j. Let $\mathbf{u}$ be $\{u_{ij}\}$, vector of polygenic effects. $\mathbf{u} \sim N(0, \sigma_a^2 \mathbf{A})$, where σ_a^2 is the additive genetic variance and $\mathbf{A}$ is additive relationship matrix. e_{ij} is residual effect. Let $\mathbf{r}$ be $\{e_{ij}\}$, vector of residual effects. $\mathbf{r} \sim N(0, \sigma_e^2 \mathbf{I})$, where σ_e^2 is error variance.

$\mathbf{A} = \{\pi_{ijk}\}$, where π_{ijk} is the coefficient of co-ancestry, which is defined as the probability that a randomly drawn gene from individual j is IBD with a randomly drawn gene from individual k at one locus. Bernardo (2010) provides typical values for various relationships in crop species. Specifically, $\pi_{ijk} = 1/2$ for full sibs if parental lines are full inbred and unrelated and 1/4 for half sibs if the common parental line is fully inbred and all the parental lines (the common parental line and other parental lines) are unrelated. Note that the off-diagonal components of $\mathbf{A}$ multiplied by σ_a^2 model the resemblance (covariance) of individuals due to IBD of segregating QTLs not due to a common environment and non-segregating QTLs within families.

Calculation of $\mathbf{A}$ requires a reference population which usually is the base of an observed pedigree (Lynch and Walsh 1998). For example, the reference could be considered as the parental generation of a FBAM population. Sometimes, crosses of an FBAM population may be connected by some common parental lines, e.g., the NAM mating design and NCD-1. Two options are available to calculate $\mathbf{A}$. One is that the coefficient of co-ancestry of offspring individuals from two different crosses is also calculated. Another is to calculate the coefficient of offspring individuals within crosses only and ignore their relationships across crosses with common parental lines, i.e., the coefficients of co-ancestry are assumed to be zero for pairs of offspring individuals from different crosses. Invoking the latter option is reasonable because IBD of individuals from different crosses can be estimated via family mean effects μ_i.

We propose a modified model:

$$y_{ij} = \mu + \beta g_{ij} + \mu_i + u_{ij} + e_{ij}$$

where β is regression coefficient (the genetic effect of marker being tested) and the others are described as above. Compared with the former model, this μ_i contains $\beta_b b_i$. Inclusion of such family effects as random variables provides a flexible way to correct for family mean effects and may increase the power of QTL detection especially in case of a large number of families with small numbers of progeny per family (Valdar et al. 2009). However, family mean effects μ_i could be modeled as fixed for fast computation in case of large numbers of progeny per family. In this way, only the within family information are used to detect QTLs.

Very often, multi-location phenotyping is conducted in crop species. For this purpose, we propose:

$$y_{\ell ij} = \mu + \beta_\ell g_{ij} + E_\ell + \mu_{\ell i} + u_{\ell ij} + e_{\ell ij}$$

where μ and g_{ij} are described as above. β_ℓ is the genetic effect of the locus being tested at location ℓ and is modeled as a fixed effect. E_ℓ is location effect, $\ell = 1, \ldots, L$, and is also modeled as a fixed effect. $u_{\ell ij}$ is a parameter representing polygenic effects of individual i of family j at location ℓ, including main background genetic effects and the background genetic effects by environment interaction. Let $\mathbf{u} = \{u_{\ell ij}\}$ represent a vector of background genetic effects, and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$. $\mathbf{G} = \sum_{\mathbf{g}} \otimes \mathbf{A}$, where $\sum_{\mathbf{g}}$ is covariance matrix of background genetic effects across locations, $\mathbf{A}$ is described as above, $\otimes$ represents direct product operation. $\sum_{\mathbf{g}}$ can be modeled by factor analytic model (Burgueno et al. 2012; So and Edwards 2011) or as a compound symmetry covariance matrix (So and Edwards 2011). Let $\mathbf{s} = \{\mu_{\ell i}\}$ represent a vector of family effects, $\mu_{\ell i}$ at location ℓ, and $\mathbf{s} \sim N(0, \mathbf{S})$. $\mathbf{S} = \sum_{\mathbf{s}} \otimes \mathbf{I}$. Similarly, $\sum_{\mathbf{s}}$ can be modeled by factor analytic model or as a compound symmetry covariance matrix. Another option is that $\mu_{\ell i}$ can be modeled as a fixed effect for fast computation especially when large family sizes are used. Let

$\mathbf{r} = \{e_{\ell ij}\}$ be a vector of residual effects, and $\mathbf{r} = \mathrm{N}(\mathbf{0}, \mathbf{R})$. $\mathbf{R} = \mathrm{daig}\ (\sigma^2_{\mathrm{e}\ell},\ \ell = 1, \ldots, L) \otimes \mathbf{I}$ or $\sigma^2_{\mathrm{e}}\mathbf{I}$.

The above models are typical mixed linear models and are easily specified and implemented in commercial software, such as Proc Mixed of SAS (SAS Institute 2004) and ASReml (VSN international 2010).

NAM stepwise regression

This method (Yu et al. 2008) was first applied to a large scale FBAM population. The relationship between phenotypic values and markers is modeled as:

$$y_{ij} = \mu + \mu_i + \sum_{\mathrm{m}} \beta_{\mathrm{m}} x_{ij\mathrm{m}} + e_{ij}$$

where y_{ij} is the phenotypic value of individual j in family i, μ is the grand mean, μ_i is the family mean effect for family i, $x_{ij\mathrm{m}}$ is the genetic score of individual j in family i at marker locus m, β_{m} is regression coefficient of marker locus m and e_{ij} is residual. Note that the family mean effect is modeled as fixed and the within family information only is used for detecting QTLs. A standard stepwise selection algorithm is used to identify the markers included in the model to represent QTLs. Yu et al. (2008) indicated that there was no difference between models with and without parameters representing family effect μ_i. We speculate that their result was due to the fact that the number of progeny (200) per family is much larger than the number of families (25) in the population.

Modified stepwise regression

Guo et al. 2010 used the following model:

$$y_{ij} = \mu + \mu_i + \beta x_{ij} + \sum_{ci} \delta_{ci} z_{cij} + e_{ij}$$

where y_{ij} is the observed trait value of segregating line j in family i, and μ_i is the mean effect of family i. x_{ij} is the imputed genetic score of segregating line j in family i at a marker locus being tested. z_{cij} is the genetic score of segregating line j for cofactor marker c_i linked to a QTL in family i. δ_{ci} is regression coefficient of cofactor marker c_i in family i. Originally, the above model was developed for data in which a targeted genomic region is genotyped with high density markers but sparse genotyping is conducted on non-targeted regions. As noted, however, genome-wide dense genotyping data could be available for a FBAM population. Therefore, we propose a modified model:

$$y_{ij} = \mu + \mu_i + \beta x_{ij} + \sum_{c} \delta_c z_c + e_{ij}$$

where z_c is the genetic score of cofactor marker c associated with a QTL and δ_c is a regression coefficient. In contrast to the former model, cofactor markers are created which are applied to the whole FBAM population not a specific family.

Cofactor markers linked to QTLs are first identified with an initial scan of the genome. Originally, cofactor markers are separately identified for individual families using composite interval mapping. When genome-wide dense genotyping data is available, cofactor markers can be identified in a whole FBAM population instead of individual families using the stepwise regression. Identification of cofactor markers can be separately conducted for independent linkage groups, because markers on independent linkage groups are reasonably assumed to be independent in a FBAM population.

After cofactor markers have been identified, all marker loci on a linkage group are tested for significant associations with a trait conditional on cofactor markers of other linkage groups. The marker with maximum test statistic or the minimal p value for the specific linkage group is declared to be a functional marker (FM) if its test statistic or p value exceeds a pre-determined threshold value. If a FM is identified, additional FMs are searched using a modified stepwise regression procedure. The algorithm consists of an adding step and an updating step. In the adding step, all marker loci are tested conditional on the already detected FM(s) plus the cofactor markers of the other linkage groups. In the updating step, previously identified FMs are re-evaluated as each FM is sequentially excluded while all other FMs and the cofactor markers of other linkage groups are included in the model as covariate(s). This updating step is repeated for each of the previously identified FMs. The adding and updating steps are repeated until no further significant associations are identified. FMs are separately searched by linkage groups. The modified procedure is superior over forward selection when multiple FMs exist in a genomic region or a linkage group (Guo et al. 2010).

Forward regression with bootstrap sampling

The model was described by Valdar et al. (2009), with some notational translation for consistency with the models described herein, as:

$$y_i = \mu + \sum_{\mathrm{m}} \beta_{\mathrm{m}} x_{i\mathrm{m}} + e_i$$

where y_i is the phenotypic value of individual i and $x_{i\mathrm{m}}$ is the genetic score of individual i at marker locus m. In the original model, a set of known covariates such as environmental covariates are included and $x_{i\mathrm{m}}$ could be a haplotype locus instead of a single marker locus. A forward selection algorithm is applied to resampled data sets produced by non-parametric bootstrapping and subsampling.

They used model-average statistics to calculate the probability of loci being included in the model selection. It is expected that this method can provide robust QTLs. Tian et al. (2011) and Kump et al. (2011) applied this method to the maize NAM population.

Haplotype interval mapping

This method was developed for testing haplotypes consisting of several markers instead of a single marker locus (Jansen et al. 2003).

The basic statistical model is described as:

$$y_{ij} = \mu_i + \beta_{h1} x_{1ij} + \beta_{h2} x_{2ij} + e_{ij}$$

where μ_i is family mean effect, x_{1ij} and x_{2ij} are separately haplotypes h1 and h2 which an offspring individual j carries in family i, β_{h1} and β_{h2} are the effects of haplotype h1 and h2, respectively. The ancestral genome blocks in the parents of the crosses are identified via haplotype analysis of parental lines and haplotypes at one block are then defined as alternative alleles and their effects are parameterized. Another optional model was also developed in which cofactor markers are added to control background QTLs (Jansen et al. 2003). Note that different blocks of loci may have different numbers of haplotypes and therefore different loci may have test statistics with different numbers of degrees of freedoms. A solution is to use p values, which can be transformed from test statistics.

Rare alleles and next generation sequencing technologies

In humans, it is recognized that less frequent or rare alleles, with large effects, are likely more important than postulated by the well-known common disease-common variant models of complex traits (Cirulli and Goldstein 2010). Although many geneticists believe that the causal variants underlying complex traits in humans may have regulatory roles due to their small effects, another line of evidence indicates that association signals credited to common variants could be created by multiple rare variants in the same genomic region, also referred to as synthetic associations (Dickson et al. 2010), or may represent a less frequent allele with larger genetic effects (International HapMap Consortium 2003). Several rare alleles have been identified in common human diseases (Stanklewicz and Lupski 2010; Rival et al. 2011; Kiezun et al. 2012) and in plant species (Yan et al. 2010). These less frequent or rare alleles could be functional variants, such as non-synonymous, nonsense or splice variants. In fact, rare alleles of large effects may play roles in agronomic traits in crop species. For example, several major effect genes have been identified for plant height in wheat and of them, *Rht1* and *Rht2* were successfully used in breeding to bring a "green revolution" (Gale and Youssefian 1985).

Next generation sequencing technologies (Metzker 2010) provide a powerful tool for genotyping all potential genomic variants. FBAM can be used as a cost-effective way to apply these technologies in crop species. The strategy is to sequence all the parental lines involved in a FBAM population and then impute scores onto the segregating progeny. However, sequencing all the parental lines may still be expensive especially when a large number of parental lines are involved. Few studies are available in application of next generation sequencing technologies in plant species (Huang et al. 2010; Lam et al. 2010). If sequence data is available, a potential strategy is to use existing data. For example, Huang et al. (2010) sequenced 517 diverse rice landraces and identified ~3.6 million SNPs using next generation sequencing technologies. They also performed association mapping analysis for 14 agronomic traits. However, this association panel may still have limited power to detect rare alleles. Therefore, lines of contrasting phenotypes could be selected for generation of a FBAM population. The population and their parental lines could be genotyped using a sparse density of markers, such as 1 marker every 5 cM, and then 3.6 million SNPs of genotypes can be imputed onto the segregating progeny of the FBAM population.

An alternative strategy is to deep-sequence regions with known associations between segregating genotypes and phenotypes. Rival et al. (2011) sequenced 56 genes from the regions associated with Crohn's disease in 350 cases and 350 controls and then identified a dozen novel alleles from identified 70 rare and low frequency allelic variants in a larger size population consisting of 9 independent case–control panels. An analogous FBAM approach in crop species is to deep-sequence regions known to have QTLs from prior studies in parental lines and then detect allelic variants in a FBAM population. A FBAM population is directly genotyped or imputed onto for the identified variants. Associated regions could come from: (1) previously reported QTL regions. A large number of QTLs have been identified in various traits in crop species. (2) QTL regions detected using the current FBAM population through multiple family QTL mapping analysis (Xu 1998; Li et al. 2005; Guo et al. 2006) when FBAM are genotyped using a sparse density of markers, (3) QTL regions detected using the above FBAM data analysis methods when the involved parental lines or the current FBAM population are genotyped using a high density of molecular markers.

A final strategy to consider consists of sequencing selected diverse parental lines representing extreme phenotypes. Cirulli and Goldstein (2010) developed an extreme trait sequencing strategy for human association

mapping. It includes whole genome-sequencing of a small number of individuals with extreme phenotypes and genotyping a larger sample using the identified variants. An analogous version for FBAM in plant species is that a small number of lines with extreme phenotypes are sequenced and associations are detected in a FBAM population which is genotyped or imputed onto for the identified variants.

In summary, FBAM will provide a powerful and cost-effective way to identify the causal variants underlying complex traits especially through application of next generation sequencing technologies. A variety of methods can be used for different situations. However, a FBAM population may be divided into groups and evaluated for phenotypes in multiple locations due to adaptation of plant species or evaluation of a FABM population may be needed for more information at multiple locations or years. The methods discussed above assume that a FBAM population is evaluated for phenotype data in a relatively uniform environment although we propose a linear mixed model approach for multi-environmental phenotyping through extending human QTDT. Epistasis may be important for crop species and a FBAM population may be a powerful source due to its involvement of diverse parental lines. But no available methods handle this issue.

Acknowledgments The authors thank Drs. Gilles Gay and Robert Bensen, Syngenta biotechnology, Inc, for critical reading of the manuscript.

Appendix: Conditional probabilities of QTL genotypes given flanking markers in DH_{F2} populations

In maize, DH_{F2} can be produced through inducing haploids from F_2 plants using inducer lines and doubling them using chemical colchicine (Chase 1951; Bordes et al. 1997; Bernardo 2009). Consider three linked loci A, Q and B, which are fixed with AAQQBB and aaqqbb for two parental lines, respectively. r_1 is recombination rate between A and Q loci, r_2 between Q and B loci and r between A and B loci. Due to the high density of markers we assume that probability of double-cross over events per meiosis is zero. Therefore, $r = r_1 + r_2$.

Production of DH_{F2} involves three processes. Firstly, F_1 plants (AQB//aqb) produce a total of six gametes: (1) No recombination event. AQB and aqb, with a frequency of $(1 - r)/2$, respectively. (2) A single recombination event between A and Q. aQB and Aqb, with a frequency of $r_1/2$, respectively. (3) A single recombination between Q and B. AQb and aqB, with a frequency of $r_2/2$, respectively. Secondly, female and male gametes are randomly mated to produce a total of 21 F_2 phased genotypes: (1) No recombination event. AQB//AQB, aqb//aqb and AQB//aqb. (2) A recombination event in one gamete between A and Q. aQB//AQB, aQB//aqb, Aqb//AQB and Aqb//aqb. (3) A recombination event in one gamete between Q and B. AQb//AQB, AQb//aqb, aqB//AQB and aqB//aqb. (4) Recombination events in both gametes between A and Q. aQB//aQB, Aqb//Aqb and aQB//Aqb. (5) Recombination events in both gametes between Q and B. AQb//AQb, aqB//aqB, AQb//aqB. (6) Recombination events in both gametes, one occurs between A and Q, but the other occurs between Q and B. AQb//aQB, aqB//aQB, AQb//Aqb and aqB//Aqb. The frequency of one phased genotype is two times the product of two F_1 gamete frequencies if the two gametes have different genotypes and the product of two F_1 gamete frequencies if the two gametes have the same genotype. For example, AQB//AQB is $(1 - r)^2/4$ and aQB//AQB is $r_1(1 - r)/2$. F_2 plants produce haploids which are equivalent to gametes. For each phased F_2 genotype, the frequencies of gametes are produced as with F_1 described above and then be multiplied by its F_2 genotype frequency to obtain the frequencies of corresponding double haplotype frequencies. For example, an F_2 genotype of aQB//Aqb, which has a frequency of $(r_1)^2/2$, produces haploids: aQB with a frequency of $(1 - r)/2$, Aqb with a frequency of $(1 - r)/2$, AQB with a frequency of $r_1/2$, aqb with a frequency of $r_1/2$, aQb with a frequency of $r_2/2$ and AqB with a frequency of $r_2/2$. The frequencies of haploids aQB, Aqb, AQB, aqb, aQb and Aqb are $r_1^2(1 - r)/4$, $r_1^2(1 - r)/4$, $r_1^3/4$, $r_1^3/4$, $r_2^3/4$, and $r_2^3/4$, respectively. A total of eight double haploid genotypes are produced from F_2 plants: AAQQBB, aaqqbb, AAqqBB, aaQQbb, aaQQBB, AAqqbb, aaqqBB and AAQQbb. Their frequencies are as follows:

$$
\begin{aligned}
P(\mathrm{AAQQBB}) = P(\mathrm{aaqqbb}) &= p_1(2 - r) + p_2(1 - r_2/2) \\
&\quad + p_3(1 - r_1/2) + p_4 r_1 + p_5 r_2 + 0.5 p_6 r \\
P(\mathrm{AAqqBB}) = P(\mathrm{aaQQbb}) &= 0.5 p_2 r_2 + 0.5 p_3 r_1 \\
&\quad + p_4 r_2 + p_5 r_1 + 0.5 p_6 r \\
P(\mathrm{aaQQBB}) = P(\mathrm{AAqqbb}) &= p_1 r_1 + p_2(1 - r_2/2) \\
&\quad + 0.5 p_3 r_1 + p_4(2 - r) + p_6(1 - r/2) \\
P(\mathrm{aaqqBB}) = P(\mathrm{AAQQbb}) &= p_1 r_2 + 0.5 p_2 r_2 \\
&\quad + p_3(1 - r_1/2) + p_5(2 - r) + p_6(1 - r/2)
\end{aligned}
$$

where $p_1 = (1 - r)^2/4, p_2 = r_1(1 - r)/2, p_3 = r_2(1 - r)/2, p_4 = r_1^2/4, p_5 = r_2^2/4$ and $p_6 = r_1 r_2/2$.

On the ends of chromosomes a single marker could be used to impute the QTL genotypes. Assume that Q locus is linked with marker locus with recombination rate of r between them. The frequencies of genotypes in DH_{F2} are:

$$P(\mathrm{AAQQ}) = P(\mathrm{aaqq}) = p_1(2-r) + p_2 + p_3 r$$
$$P(\mathrm{AAqq}) = P(\mathrm{aaQQ}) = p_1 r + p_2 + p_3(2-r)$$
$$\text{where } p_1 = (1-r)^2/4, p_2 = r(1-r)/2 \text{ and } p_3 = r^2/4$$

References

Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292

Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling based tests of linkage and association for quantitative traits. Am J Hum Genet 64:1754–1764

Anderson L, Georges M (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genetics 5:202–212

Bernardo R (2009) Should maize doubled haploids be induced among F_1 or F_2 plants. Theor Appl Genet 119:255–262

Bernardo R (2010) Breeding for quantitative traits in plants. Stemma Press, Minnesota

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor Appl Genet 113:206–224

Bordes J, Dumas de Vaul R, Lapierre A, Pollacsek M (1997) Haplodiploidization of maize (*Zea mays* L.) through induced gynogenesis assisted by glossy markers and its use in breeding. Agronomie 17:291–297

Buckler E, Gore M (2007) An Arabidopsis haplotype map takes root. Nat Genet 39:1056–1057

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C et al (2009) The genetic architecture of maize flowering time. Science 325:714–718

Burdick JT, Chen W, Abecasis GR, Cheung VG (2006) In silico method for inferring genotypes in pedigrees. Nat Genet 38:1002–1004

Burgueno J, Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719

Chase SS (1951) Production of homozygous diploids of maize from monoploids. Agron J 44:263–267

Churchill GA et al (2004) The collaborative cross: a community resource for the genetic analysis of complex traits. Nat Genet 36d:1133–1137

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole genome sequencing. Nat Rev Genet 11:415–425

Darvasi A, Weinreb A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943–951

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome wide associations. PLoS Biol 6:e1000294

Farnir F, Grisart B, Coppieters W, Riquet J, Berzi P, Cambisano N, Karim L et al (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred hal-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161:275–287

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267

Gale MD, Youssefian S (1985) Dwarfing genes in wheat. In: Russell GE (ed) Progress in plant breeding 1. Butterworths, London, pp 1–35

George AW, Visscher PM, Haley CS (2000) Mapping quantitative trait loci in complex pedigree: a two- step variance component approach. Genetics 156:2081–2092

Guo B, Beavis WD (2011) In silico genotyping of the maize nested association mapping population. Mol Breed 27:107–113

Guo B, Sleper DA, Sun J, Nguyen HT, Arelli PR, Shannon JG (2006) Pooled analysis of data from multiple quantitative trait locus mapping populations. Theor Appl Genet 113:39–48

Guo B, Sleper DA, Beavis WD (2010) Nested association mapping for identification of functional markers. Genetics 186:373–383

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 89:315–324

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) Family based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61–69

Huang X, Wei X, Sang T, Zhao Q, Feng Q et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

International Hapmap Consortium (2003) The international HapMap project. Nature 426:789–796

Jansen RC, Jannink JL, Beavis WD (2003) Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. Crop Sci 43:829–834

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM et al (2012) Exome sequencing and the genetic basis of complex traits. Nat Genet 44:623–630

Kingsmore SF, Lindquist IE, Mudge J, Gesler DD, Beavis WD (2008) Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov 7:221–230

Kump KL, Bradbury PJ, Buckler ES, Belcher AR, Oropeza-Rosas M et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43:163–168

Laird NM, Lange C (2006) Family based designs in the age of large scale gene association studies. Nat Rev Genetics 7:385–394

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SM, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Lander ES, Bostein (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lange C, DeMeo DL, Laird NM (2002) Power and design considerations for a general class of family based association tests: quantitative traits. Am J Hum Genet 71:1330–1341

Lee SH, Werf JHJ (2004) The efficiency of designs for fine mapping of quantitative trait loci using combined linkage disequilibrium and linkage. Genet Sel Evol 36:145–161

Li R, Lyons MA, Wittenburg H, Paigen B, Churchill GA (2005) Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics 169:1699–1709

Lund MS, Sorensen P, Guldbrandtsen B, Sorensen DA (2003) Multi fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. Genetics 163:405–410

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Inc, Sunderland

Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67:146–154

Metzker M (2010) Sequencing technologies-the next generation. Nat Rev Genetics 11:31–46

Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379

Monks SA, Kaplan NL (2000) Removing the sampling restrictions from family based tests of association for a quantitative trait locus. Am J Hum Genet 66:576–592

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. Nat Rev Genetics 13:85–96

Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211–223

Rival MA, Beaudoin M, Gardet A, Stevens C, Sharma Y et al (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43:1066–1073

SAS Institute (2004) SAS/STAT 9.1 user's guide. SAS institute, Cary, NC

Snape JW (1988) The detection and estimation of linkage using doubled haploid or single seed descent populations. Theor Appl Genet 76:125–128

So Y, Edwards J (2011) Predictive ability assessment of linear mixed models in multi-environment trials in corn. Crop Sci 51:542–552

Stanklewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455

Steen KV, McQueen MB, Herbert A, Raby B, Lyon H et al (2005) Genomic screening and replication using the same data set in family based association testing. Nat Genet 37:683–691

Stich B, Melchinger AE, Piepho H, Heckenberger M, Manurer HP, Reif JC (2006) A new test for family based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113:1121–1130

Tian F, Bradbury PJ, Brown PJ, Sun Q, Flint-Garcia S et al (2011) Genome-wide association study of maize identifies genes affecting leaf architecture. Nat Genet 43:159–162

Valdar W, Holmes CC, Mott R, Flint J (2009) Mapping in structured populations by resample model averaging. Genetics 182:1263–1277

VSN international (2010) ASReml 3. VSSN international Ltd., Hemel Hempstead

Xu S (1998) Mapping quantitative trait loci using multiple families of lines crosses. Genetics 148:517–524

Yan J, Kandianis CB, Harjes CE, Bai L, Kim EH et al (2010) Rare genetic variation at Zea mays crtRB1 increases β-carotene in maize grain. Nat Genet 42:322–327

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M et al (2006) A unified mixed model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551