8

Using Natural Allelic Diversity to Evaluate Gene Function

Sherry R. Whitt and Edward S. Buckler, IV

Summary

Genomics has developed a wide range of tools to identify genes that play roles in specific pathways. However, relating individual genes and alleles to agronomic traits is still quite challenging. We describe how association analysis can be used to relate natural variation at candidate genes with agronomic phenotypes. Association approaches in plants can provide very high resolution and can evaluate a wide range of alleles rapidly. We discuss issues related to experimental design, germplasm sample, molecular assay, population structure, and statistical analysis necessary for association analysis in plants.

Key Words

association analysis, candidate gene, linkage disequilibrium, LD, maize, phenotypic variation, population structure, mapping, QTL, quantitative trait loci, selection, diverse germplasm

1. Introduction

We describe a methodology for dissecting complex traits using association analysis and natural diversity. In a high diversity species such as maize, association analysis has the potential to map quantitative trait loci (QTL) with up to 5000 times better resolution than mapping with standard F_2 populations ***(1)***. In addition, association approaches may survey tens of alleles, whereas standard mapping approaches survey a maximum of two alleles. Association approaches do not require special mapping populations, but rely on the extensive history of mutation and recombination to dissect a trait. The structure of linkage disequilibrium (LD), which is the correlation between polymorphisms, and evaluation of selection is key to utilizing association analysis ***(2,3)***.

From: *Methods in Molecular Biology, vol. 236: Plant Functional Genomics: Methods and Protocols*
Edited by: E. Grotewold © Humana Press, Inc., Totowa, NJ

The use of extant natural diversity provides advantages in resolution and breadth of survey, but can also present added difficulties in accurately assessing the true cause of an association. The most serious false positives can result when unlinked markers produce a positive association because of underlying population structure. The complex breeding history of most crops and the limited gene flow in most wild plants creates population stratification within the germplasm ***(4)***.

In recent years, a few statistical methods have been developed that use independent marker loci as a means of detecting and correcting for population structure ***(5,6)*****.** These methods work on the assumption that population structure should affect all loci in a similar manner. Reich and Goldstein ***(6)*** propose scoring a moderate number of unlinked genetic markers (e.g., single nucleotide polymorphisms [SNPs] or simple sequence repeats [SSRs]) and then comparing the strength of the candidate gene association with those of the unlinked markers. We have utilized a modified approach designed by Pritchard et al. ***(7,8)***, which incorporates a test statistic of likelihood ratios that includes estimates of subpopulation allele frequencies and evaluates quantitative traits ***(1)***.

A standard procedure for carrying out association analysis on candidate genes is as follows (*see* **Fig. 1**):

1. Select positional candidate genes using existing QTL and positional cloning studies.
2. Choose germplasm that will capture the bulk of diversity present. When possible, inbred lines should be used.
3. Score phenotypic traits in replicated trials.
4. Amplify and sequence candidate genes.
5. Manipulate sequence into valid alignments and identify polymorphisms.
6. Obtain diversity estimates and evaluate patterns of selection.
7. Statistically evaluate associations between genotypes and phenotypes taking population structure into account.

2. Materials

2.1. Germplasm

Sample at least 100 inbred lines of germplasm (for a maize example, see [http://www.maizegenetics.net]). For high resolution do not choose closely related samples. In order to test for selection in a crop, collect one sample from a sister taxon to function as an outgroup for the Hudson, Kreitman, and Aguade (HKA) test ***(9)***.

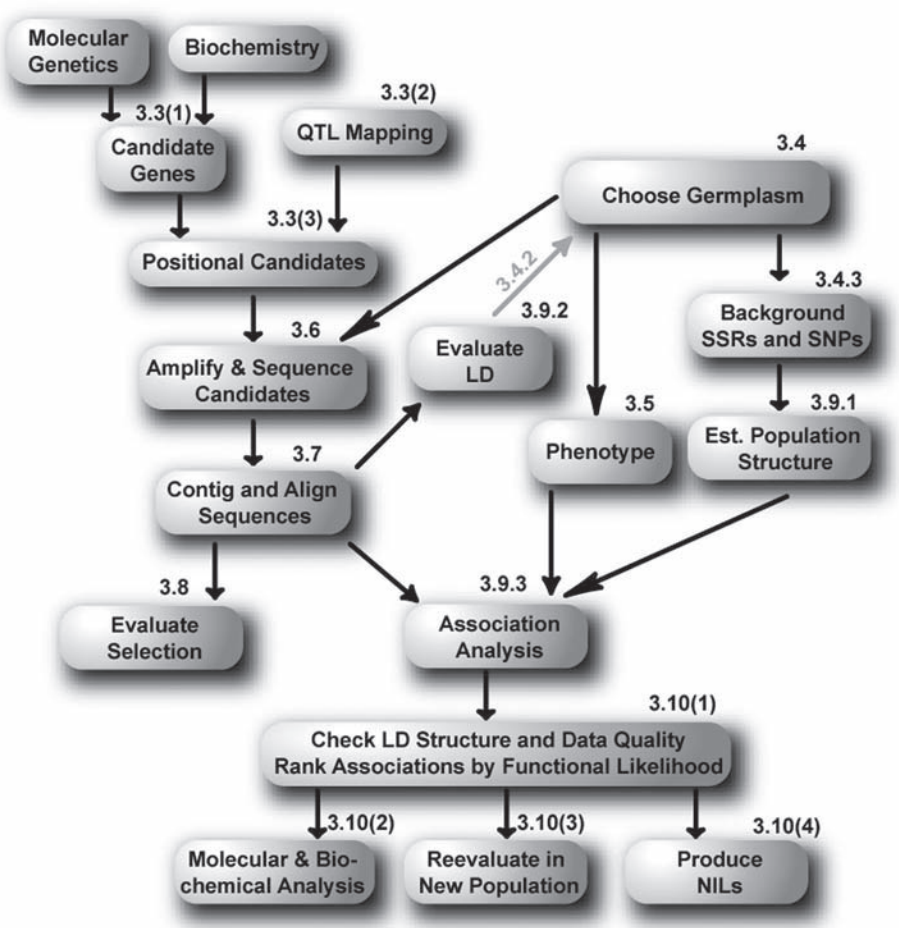

Fig. 1. Association study employs techniques from molecular biology, field sampling–breeding, bioinformatics, and statistics. The steps necessary to associate a particular genotype with a phenotypic trait are illustrated. Above each step is a numeric reference to the relevant text. The gray arrow linking "Evaluate LD" and "Choose Germplasm" signifies the potential need to revise the choice of germplasm once the structure of LD is known.

2.2. Primer Design

Primer3.0 (http://www-genome.wi.mit.edu) and PCR-Overlap (http://droog.mbt.washington.edu) provide oligonucleotide design for Linux or Unix® operating systems.

2.3. PCR

1. FailSafe™ system (Epicentre): 2× premixtures labeled A–L; contain dinucleotide triphosphates (dNTPs), Tris-based buffering solution, varying amounts of $MgCl_2$, and betaine (*see* **Note 1**).
2. FailSafe enzyme: 2.5 U/μL, a mixture of polymerases with proofreading capabilities (*see* **Note 2**).
3. Genomic DNA (33 ng/μL) (purified with DNeasy™ plant maxi kits [Qiagen]).
4. Primers.
5. QIAquick™ 8 PCR purification kit (Qiagen) (*see* **Note 3**).

2.4. Cycle Sequence

1. BigDye™ chemistry (Applied Biosystems) (we dilute enzyme with dilution buffer for quarter reactions). Dilution buffer (halfTERM dye terminating sequence [Sigma] or Half-Dye™ mixture [Denville Scientific]) (see **Note 4**).
2. High-performance liquid chromatography (HPLC) water.
3. Purified PCR template (*see* **Subheading 3.6.6.**).
4. Primers (*see* **Subheading 3.6.1.**).
5. DyeEx™ terminator removal kit (Qiagen).

2.5. Sequence Manipulation Software

1. PHRED and PHRAP versions from CodonCode (http://www.codoncode.com/) are used to assess sequence quality and contig (join) sequences ***(10)***.
2. Biolign (Tom Hall [http://www.mbio.ncsu.edu/BioEdit/bioedit.html]) is used to edit multiple alignments of contigs and evaluate SNPs. Biolign is a custom software package. MegAlign from DNASTAR and Sequencher™ from GeneCodes (http://www.genecodes.com) offer some similar features.

2.6. Software for Testing for Selection

Statistical analyses for several tests of selection are performed with DnaSP 3.0 (http://www.bio.ub.es/~julio/DnaSP.html), which has a user-friendly interface ***(11)***, and SITES (http://lifesci.rutgers.edu/~heylab) ***(12)***.

2.7. Association Analysis Software

1. Population structure software: STRUCTURE (http://pritch.bsd.uchicago.edu) is an excellent program to estimate population structure ***(7)***.
2. LD software: Arlequin (http://lgb.unige.ch/arlequin) can handle a wide range of markers and sequences ***(13)***. It can also calculate LD from genotypic data. DnaSP (http://www.bio.ub.es/~julio/DnaSP.html) manages numerous DNA sequences and can plot LD ***(11)***. PowerMarker (http://www.powermarker.net) can incorporate a wide range of markers and genotypic data and can produce plots of LD. TASSEL (http://www.maizegenetics.net) has the capability to cope with a wide range of markers, sequences, and plot LD.
3. Association software: SAS (http://www.sas.com) is a general purpose statistical

software package and can carry out a wide range of statistics useful for association analysis. STRAT (http://pritch.bsd.uchicago.edu) can be used for testing association of binary traits across structured populations ***(8)***. TASSEL (http://www.maizegenetics.net) can perform analysis of variance (ANOVA) and logistic regression association tests that control for population structure.

3. Methods

3.1. Polymerase Chain Reaction

1. Combine the following, scaling vol for number of reactions desired, to produce 25 μL total vol reactions (add genomic DNA subsequently): 12.5 μl 2× premixture, 7.0–9.0 μL HPLC-grade water, 1.0 μL-20 μ*M* forward primer, 1.0 μL 20 μ*M* reverse primer, 0.5–1.0 μL *Taq* DNA polymerase, and 1.0–3.0 μL genomic DNA.
2. Amplify the above reaction in a thermal cycler.
3. Purify PCR product using QIAquick 8 PCR purification kit (*see* **Note 3**).

3.2. Cycle Sequence

1. Prepare a standard reaction as follows in 10 μL total vol: 2 μL terminator enzyme, 2 μL dilution buffer, 4 μL PCR template, and 2 μL 3–5 μM primer.
2. Sequence reactions are cleaned via DyeEx terminator removal kit (for cycle profile *see* **Subheading 3.6.6.**).

3.3. Selection of Positional Candidate Genes

Choosing candidate genes, i.e., those genes most likely to contain the polymorphism responsible for the phenotype, is one of the most critical steps in conducting association analysis. However, candidate gene selection is currently as much an art as a science. We refer to candidate genes that fall within QTL intervals as positional candidate genes. QTL mapping often has limited resolution, but is an excellent way to narrow the search for candidates to specific chromosomal regions. Focusing on positional candidate genes will maximize the opportunity to find associations. The major aspects of choosing genes are:

1. Collect a list of genes that affect the phenotype of interest. Mutagenesis, biochemistry, various profiling technologies, comparative genomics, and positional cloning techniques–studies can aid in the identification of genes. Create a list of chromosomal positions for these candidate genes.
2. Collect a list of map positions of QTL for the trait of interest over all previous experiments. Various databases such as MaizeDB (http://www.agron.missouri.edu) and Gramene (http://www.gramene.org) provide a good starting point for determining these positions. A candidate gene should not be ruled out if only one or two populations have been mapped.
3. Compare the two lists and generate a list of all known genes with potential phe-

notypic effects in QTL confidence intervals. These positional candidate genes with the most neighboring QTL are most likely to have segregating variation at the locus. These positional candidate genes should be sampled first.

3.4. Choice of Germplasm

3.4.1. Phenotypic Diversity

The choice of germplasm is crucial to the discovery of useful alleles. In order to have enough statistical power to find an association, it is critical that the samples span the full range of phenotypic variation. To maximize the range of alleles tested, a genotypically diverse set of germplasm should be chosen. When available, marker or phenotypic surveys can be used to choose a subset of the germplasm that is most diverse. Software such as MSTRAT (http://www.ensam.inra.fr/gap/MSTRAT/mstratno.htm) and PowerMarker provide methods for helping to choose the germplasm. From a practical level, we found that a sample of 100 diverse inbred lines has enough statistical power to identify associations that control 10% of the phenotypic variation ***(1)***. Larger samples and/or more replications of phenotypic evaluation could be used to identify associations with smaller effects. Although geneticists are forced to do association studies with outbred populations, inbred lines provide a number of advantages to plant researchers. Inbred samples allow direct identification of haplotypes throughout the genome, generally have more consistent phenotypes than segregating populations, and provide evaluation of phenotype without the complications of dominance. In some species, core sets of germplasm have been defined and characterized, and these are excellent starting points for association studies.

3.4.2. Resolution and LD

The choice of germplasm will also determine the resolution of association approaches. Highly diverse germplasm has an extensive history of recombination, which can result in high-resolution association analysis. However, high resolution will require a high marker density to identify associations. Resolution of associations is directly related to the structure of LD ***(14)***. LD is the correlation between pairs of polymorphisms. One simple way to estimate LD between pairs of sites is to calculate r^2 ***(15)***. The average distance between polymorphisms, at which r^2 drops below 0.1, is a rough estimate of the resolution within a specific population. The rate of LD decay in most cases needs to be determined empirically for any given population ***(2,16)*** (*see* **Note 5**). However, the rate of LD decay may also be locus-specific, as differences in recombination rate, mutation rate, and selection history can affect LD patterns.

3.4.3. Germplasm Population Structure

The final consideration in selecting the sample population is whether to use randomly or nonrandomly mated germplasm. Unfortunately, there is little truly randomly mated breeding germplasm available, other than a few unselected synthetic populations. Many of these randomly mated populations represent a rather narrow group of germplasm, which is likely to lower resolution and harbor only a narrow range of alleles. However, if nonrandomly mated germplasm is used, population structure needs to be controlled in the statistical analyses. In addition, the genome for each sample population should be genotyped with SSRs, SNPs, restriction fragment-length polymorphisms (RFLPs), random-amplified polymorphic DNAs (RAPDs), or amplified fragment-length polymorphisms (AFLPs) to provide an estimate of population structure (*see* **Subheading 3.9.1.**). Our experience has been that 50–150 markers generally provide good estimates of population structure. The ideal markers are either a modest number of SSRs or large numbers of SNPs, while if resources are limited, AFLP may provide a good compromise.

3.5. Phenotypes

We test for associations between polymorphisms with a wide range of agronomic and physiological phenotypic traits. The phenotypic data comes from 2 to 3 field seasons of randomized plots with 10–15 plants per row, replicated across multiple environments. The measurement of phenotypic traits needs to be a balance of simplicity in data collection, biological relevancy, and reproducibility.

3.6. Gene Amplification and DNA Sequencing

Once a candidate gene has been identified, the researcher carries out a set of standard procedures including various molecular techniques. A general guideline is as follows:

1. Design compatible primer pairs from candidate gene sequence.
2. Employ PCR to amplify the target.
3. Verify product from PCR by agarose gel electrophoresis and purify the DNA.
4. Obtain a DNA sequence product directly from the PCR product, by using commercially available labeling chemistry and enzymes (*see* **Subheading 3.6.5.**).
5. Clean up sequencing reactions to eliminate excess dNTPs, enzyme, and buffer.
6. Determine nucleotide sequence by electrophoresis (*see* **Note 6**).

3.6.1. Primer Design

1. Define a “standard” allelic sequence for primer design and future alignments (*see* **Subheading 3.7.2.**).

2. Design a series of overlapping primers, based on the standard allelic sequence, across the gene via PCR-overlap in conjunction with Primer3 (*see* **Note 7**). Typical coverage is usually in 1-kbp fragments of the gene. Primers may also be designed manually by visual inspection of sequence and apply the general rules of primer design. Generate forward and reverse primers approx 18–25 bp in length with similar melting temperature (T_m) (near 60°C) and order from one of several companies offering this service (*see* **Note 8**).
3. Resuspend lyophilized forward and reverse oligonucleotides at a standard stock concentration of 100 μ*M* in 1× Tris-ethylene diamine tetraacetic acid (EDTA) buffer and store at –80°C. Oligonucleotides should be further diluted to a working stock concentration of 20 μ*M* for PCR and 3–5 μ*M* for cycle sequencing, both of which should be stored at –20°C.

3.6.2. Optimization of PCR

1. Attempt initial PCRs by combining various primer sets and buffer conditions at an annealing temperature gradient from 50°–60°C. Utilize genomic DNA from a few representative test samples before including the entire population. Buffers containing a range of $MgCl_2$ and betaine are evaluated for optimal amplification (*see* **Note 9**).
2. A standard PCR program carried out on a thermal cycler may include: 5 min denaturation at 96°C, followed by 25–35 cycles of: 30 s denaturation at 96°C, 30 s annealing at 50°–65°C, and 30 s to 4 min extension at 70°–72°C; a final extension at 70°–72°C for 5–10 min, and hold at 4°C. A typical PCR program will take from 2–4 h (*see* **Note 10**). Annealing temperatures are set a couple of degrees below the primer melting temperatures, and extension times are delineated by the size of the expected PCR product using the 1 min/1 kbp rule.
3. When optimal buffering conditions and annealing temperatures are found, the remainder of the sample population is included along with numerous negative and positive controls in subsequent PCR (*see* **Note 11**). Increased product amount can be obtained by scaling up the total reaction vol to 50 μL.
4. Most PCR products can be directly sequenced from inbred lines, because all loci are homozygous. Some researchers may wish to compare sequence diversity between domesticated species and wild relatives (e.g., *Zea mays* ssp. *mays* and *Zea mays* ssp. *parviglumis*). Due to the heterozygous nature of wild relatives, we clone the PCR product before sequencing.

3.6.3. Agarose Gel Electrophoresis

Once PCR is completed, check the product for the correct band size and amount by agarose gel electrophoresis with ethidium bromide staining. Utilize a mass ladder as a standard to determine size and quantity of product fragments. Add 6× loading dye to the mass ladder and samples prior to electrophoresing the samples at an appropriate voltage. Visualize the products by UV transillumination (*see* **Note 12**). When more than one product is obtained, the correct fragment may be excised from the agarose gel with a sterile blade.

3.6.4. Purification of PCR Product

PCR product will yield quality sequence data only when all enzyme, primer, and other reagents are removed from the reaction. Final products are purified with a vacuum manifold in an 8-strip format. The procedure takes approx 20 min for 48 samples. Product is eluted into a 96-well plate and maintained at –20°C until the template is sequenced.

3.6.5. Cloning Gene Fragments

TOPO® TA cloning kit (Invitrogen) provides superior success rates. The PCR product is ligated into an approx 4-kbp vector with thymidine overhangs and transformed into chemically competent *Escherichia coli* One Shot® TOP10® (which provides both ampicillin and kanamycin resistance, as well as blue–white colony screening). To prepare small quantities of DNA, we use Qiagen's QIAprep® 8 turbo miniprep kit. TOPO TA cloning and transformation takes approx 2 h and overnight growth in an incubator at 37°C. The miniprep preparation by vacuum manifold takes 30 min to complete. Ultimately, clones are stored as glycerol stocks in 96-well format at –80°C.

3.6.6. Cycle Sequencing

Sequence reactions should follow standard protocols for the chemistry and, in general, take approx 2 1/2 h to complete.

1. The sequence reaction is set up in a 10-µL total vol (*see* **Subheading 3.2.**).
2. The amount of template used should equate to approx 30–50 ng of DNA.
3. A typical cycle sequence program is: 92°C for 30 s, 50°C for 30 s, and 60°C for 4 min, repeated 25 times, then maintain at 4°C until removal from thermal cycler.
4. Sequence product is cleaned up to eliminate excess enzyme and primer. Prepackaged kits are available with a procedure that takes approx 20 min (*see* **Subheading 3.2.**).
5. DNA Sequence is obtained by Model 3700 capillary electrophoresis (Applied Biosystems) in a 96-well format. Sequence products are light-sensitive. Keep exposure to a minimum.

3.7. Sequence Manipulation

Sequence manipulation involves database handling of trace files, applying quality scores to individual bases, and contiging (joining) and aligning sequence data.

3.7.1. Join Sequence Fragments

DNA sequence is received as trace files of the chromatograms and text files of the nucleotide sequence. The trace files are sorted by gene and sample to

individual gene folders, accordingly. Quality scores and contigs are obtained using the CodonCode versions of PHRED and PHRAP (*see* **Note 13**). The quality scores are reported in spreadsheet format as the total number of bases with a Phred score of 20 or higher for a particular sample. Sequences with more than 400 bases with scores of 20 and higher are included in the alignment. CROSSMATCH is used to remove vector sequence if fragments were cloned. Phrap contigs the sequences to produce an ".ace" file, which contains nucleotide reads and associated Phred scores.

3.7.2. Align Sequences

.ace files are aligned in the software program Biolign. When possible, a published sequence is used as the standard in a framework around which the alignment is built. The standard sequence can be used to delineate base call differences in any of the sequenced samples. Phrap quality coloring indicates the quality of sequence, by highlighting specific bases with different colors based on phred quality scores <30. Regions with low quality (e.g., <Phred 20) are converted to the missing data symbol "?" or "N." Prior to calculating diversity indices, introns and exons are delimited, and then the annotated contiged sequence is saved in the NEXUS file format.

3.8. Evaluation of Diversity and Selection

The power to detect associations depends on genotyping, genetic architecture, and accurate phenotypic evaluations. If there are complications with any of these three factors, there may be little statistical power in relating a gene to a specific phenotype. However, the signature of artificial selection can also be used to provide evidence that a specific gene is important for controlling phenotypic variation. If a gene has been a target of selection through the domestication and breeding process, then it is likely to control an agronomic phenotype and could be useful in future breeding and genetic manipulation. Nucleotide diversity surveys can powerfully detect several forms of selection. We describe two tests of selection that can be useful in finding genes that play key roles in phenotypic variation. The Tajima's D test evaluates diversity within a species to find evidence of selection, while the HKA test compares nucleotide diversity within a species to the nucleotide difference with a related species. Diversity and selection estimates can be obtained as follows:

1. DnaSP enables the user to calculate several indices of genetic diversity, divergence, and selection. Generate an aligned nucleotide sequence with codon assignment in NEXUS file format (*see* **Note 14**).
2. Calculate diversity measures for the sequence data. We report π (the average number of nucleotide differences per site between two sequences) and θ (similar

to π, but focuses on the number of segregating sites) for nonsynonymous and synonymous sites and LD (*see* **Subheading 3.9.2.**).

3. Perform tests of selection. Tajima's D test statistic compares diversity based on average number of differences (π) vs the number of segregating sites (θ), hypothesizing all mutations are selectively neutral ***(17)***. The statistic may also reflect demographic changes or population structure, so caution is needed in interpreting these results.
4. The HKA test examines the ratio of intraspecific diversity to interspecific divergence using an outgroup ***(9)*****.** The outgroup species should have diverged just before the time when the alleles within the target species began diverging (*see* **Note 15**). Within the DnaSP program, a second data window must be opened containing the outgroup sequence. The test is calculated by comparing silent θ and silent K (divergence). A low value relative to other loci suggests that selection, specifically, has reduced diversity at a particular locus. Neutral loci are needed for comparison in this test.
5. A significant selection test may mean little molecular variation with which to find associations. These tests indicate that selection has occurred, but they are generally ambiguous as to why selection has occurred.

3.9. Statistical Applications to Find Genotype–Phenotype Associations

3.9.1. Estimate Population Structure

If the samples are not randomly mated, it is critical that population structure be included in the association analysis. The STRUCTURE software is a good way to estimate population structure for association approaches.

1. Convert genotypic marker data (e.g., random SSR or SNP data throughout genome) to STRUCTURE format (*see* **Note 16**).
2. Run STRUCTURE and test with one population, continue to increase the population number until the maximum likelihood is identified. Cycles (100,000) for both burn-in (the period where the model explores the parameter space) and likelihood estimation seems to work well. At least five repetitions should be conducted for each population size (*see* **Note 17**).
3. Extract the Q matrix from the optimal result for later use (*see* **Subheading 3.9.3.**).

3.9.2. Evaluate LD

Understanding the structure of LD for a specific locus will, in turn, reveal the association resolution possible at that locus. For example, if LD decays within 1000 bp, then 1 or 2 markers per 1000 bp will be needed to identify associations.

1. DnaSP, Arlequin, or TASSEL will calculate LD between pairs of polymorphisms (r^2 or D'). Use any one of these programs to calculate all pairwise estimates of LD.

2. Plot the distance between the polymorphisms in basepairs vs LD (e.g., r^2). From this plot, one can estimate the point at which r^2 is below 0.1 (a rough estimate of the resolution of the association study).
3. Plot the strength of LD between all pairs of sites, which can be graphically done in PowerMarker or TASSEL. The graph will identify blocks of high LD and will show which sets of sites are highly correlated. Association approaches will have trouble differentiating between blocks of highly correlated sites.

3.9.3. Evaluate Associations

1. Filter polymorphisms: the segregating sites need to be extracted from the sequence alignments either by hand or by programs such as TASSEL and DnaSP. Normally, polymorphisms that are present in less than three samples or with a frequency <5% are not included in the analyses. These low frequency polymorphisms may be the product of PCR or sequencing error. Additionally, there is rarely enough statistical power to test for association at these low frequency polymorphisms. Insertions and deletions also need to be identified and coded for analysis. TASSEL does this automatically, while DnaSP ignores this type of polymorphism.
2. Randomly mated samples: when samples are truly randomly mated, no correction for population structure is required. If the trait is binary (e.g., yellow vs white kernels), then a series of chi-square tests (χ^2) can be used to evaluate whether the segregating polymorphisms associate. If the trait is quantitative, then a series of *t*-tests or ANOVA can be used to evaluate the associations (*see* **Note 18**).
3. Structured samples: when population structure is present, statistical analysis must account for it. If the trait is binary, the STRAT program can be used to evaluate the associations. If the trait is quantitative, either SAS or TASSEL can be used to implement the logistic regression ratio test. In the null hypothesis H_0, candidate polymorphisms are independent of phenotype; while in the alternative hypothesis H_1, candidate polymorphisms are associated with the phenotype. The probability of each hypothesis is compared in the following way:

$$\Lambda = \frac{\Pr_1(C;T,\hat{Q})}{Pr_0(C;\hat{Q})}$$

Where *C* is the genotype of the candidate polymorphism for all lines, and *T* is the trait value for all lines. In this test, the difference in the natural logarithm likelihoods of the model with (Pr_1) and without (Pr_0) the trait is the test statistic Λ (*see* **Note 19**). Since the distribution of Λ is not known precisely, permutations should be used to determine significance. If several sites with high LD are being scored, then the maximum Λ over all sites is used as the test statistic Λ_{max}. Permutations are calculated based on this Λ_{max} statistic.

4. Permutations to determine significance: these statistical tests will result in a *P*-value associated with each polymorphism-trait pair. However, for many associa-

tion tests, there will be 10s or 100s of polymorphisms to test. The normal modification for multiple tests would be a Bonferroni correction, however, this is far too conservative for highly correlated polymorphisms. The goal of permutations is to determine the number of independent tests, which is confounded by LD, and account for nonnormality in trait distributions. The trait values should be permuted relative to the fixed haplotypes ***(18)***, and then associations recalculated for 100–1000 permutations. Pritchard et al. ***(8)*** suggests permutations based on population structure, and this approach is implemented in STRAT and TASSEL.

5. Compare the permuted *P*-value to the distribution of *P*-values for random markers across the genome. In some cases, the estimates of population structure do not explain all of the structure. Subsequently, the random markers used for estimating population structure could be used, as could data from unrelated candidate genes. The candidate gene *P*-value could be rescaled based on the *P*-values for the random markers. For example, if the candidate gene had a *P*-value of 0.03, but 7% of the random markers had a *P*-value <0.03, then the candidate genes *P*-value could be rescaled to 0.07. This is probably a conservative test, as some of the random markers are likely to be truly associated with the trait.

3.9.4. Evaluate Associations Using TASSEL

Outlined below is a step-by-step example of how to use TASSEL to carry out association tests with structured populations. TASSEL can work with data stored in databases, but in this example, we describe TASSEL use with flat files (unclick the DB button in the main window).

1. Download and install the program by going to (http://www.maizegenetics.net).
2. Create a sequence alignment in PHYLIP format or CLUSTAL format. Many sequence editors can produce these alignments, such as CLUSTALW or BioEdit. Load the sequence alignment into TASSEL by clicking the Data button and then the Gene button and selecting your alignment file.
3. Create text files in the format described in the TASSEL help section for the population structure matrix (Q matrix from 3.7.1) and the trait data. Load the trait file by clicking on the Trait button, and the Q matrix by clicking the Pop button. It is critical that taxa names are exactly the same for the sequence alignment, Q matrix, and trait matrix.
4. Remove the invariant and low frequency sites from the sequence alignment by selecting the sequence alignment and clicking the Sites button. We normally examine sites with a minimum frequency of 0.05, as less frequent sites often have little power to detect significant results with samples less than several hundred taxa.
5. Join the filtered alignment with the Q and trait matrices by selecting all three matrices and the clicking the ∩ Join button, which will produce the intersection of these datasets.
6. Click Analysis and then Struct. Assoc. to carry out a structured association analy-

sis. Use the arrows to move the Q matrix values to the Pop Structure Estimate list. Generally at least 1000 permutations should be run.

7. These results will be summarized in two reports. The first report summarizes the results for the entire data set and accounts for the multiple tests conducted. The second report provides information on how individual sites were associated. Results may be viewed in tabular or graphical format by clicking the Results button.

3.10. Interpretation of Genotype–Phenotype Associations

Once an association is empirically determined, the validity of the association must be ascertained.

1. Which associating polymorphisms most likely control the trait? First, it is critical that genotypes be rechecked, and results should be examined to determine if phenotypic outliers are driving associations. Association studies will often find multiple polymorphisms that significantly associate. Carefully examining the LD structure surrounding the association can help identify this suite of polymorphisms and where more sampling may be needed. Although the most significant site is the most likely cause for the association, many of the slightly less significant sites could actually be the functional cause of the phenotypic variation. We find that breaking the polymorphisms into likely functional (biologically significant) vs likely silent is useful in developing lists of sites for future evaluation. Radical coding sequence changes, changes in conserved promoter motifs, changes within splicing motifs, and large insertions–deletions are generally put in the likely functional list.
2. The most straightforward way to prove an association is to evaluate the candidate polymorphisms in an entirely different population sample. Only polymorphisms that are closely linked to the cause of a phenotype should be significant in a second study. It is important that the population structure of the second sample is truly independent of the first sample. Only the candidate polymorphisms need to be retested in the new sample. In maize, we are using randomly mated synthetic populations for reevaluation of association studies.
3. In some cases, associations will suggest a molecular or biochemical mechanism of action. Following-up hypotheses generated by association analysis with molecular biology and biochemistry can be very productive, but it should be warned that association studies could be picking up on effects that only explain a few percent of the variation. Many biochemical and molecular approaches may not be quantitatively sensitive enough to detect such small changes at a molecular level.
4. Final proof of the association can be obtained through marker-assisted selection and production of near isogenic lines (NIL).

3.11. Conclusions

Mapping with F_2 or derived populations is powerful for evaluating two alleles with low resolution. In contrast, association analysis can evaluate numerous alleles at high resolution. These two approaches are complementary. The successful integration of these two approaches will allow the rapid dissection of almost any trait within a few years time. The key to association analysis is the choice of germplasm, quality of phenotypic data, and use of statistical analyses to control for population structure. The combination of association mapping and QTL mapping could make it routine to dissect complex traits down to the single gene level.

4. Notes

1. Maize is particularly GC-rich and requires additional components for optimal PCR; the FailSafe system is expensive, but allows for easier optimization as necessary reagents are premixed and contain a wide range of concentrations.
2. We have had success with *Taq* DNA polymerase (no proofreading) as well.
3. These kits provide exceptional product for subsequent cycle sequencing, but are fairly expensive. Phenol–chloroform extraction is also used as an inexpensive alternative.
4. We tried a homemade recipe, but achieved 100–200 fewer bases with sequence results. Others have had success with a solution containing Tris, $MgCl_2$ at pH 9.0, and water.
5. In maize for example, LD decays within 600 bp for landraces of maize ***(16)***, within 2000 bp for diverse breeding inbred lines ***(2)***, whereas LD persists up to 100,000 bp for elite inbred lines ***(14)***.
6. We typically obtain 500–600 bp reads on average from a Model 3700 analyzer. Alternate sequence methodologies are available, such as Model 377 and MegaBase technologies.
7. We include a library of common repetitive elements in the mispriming library, which seems to improve efficiency especially for longer amplicons.
8. We use Operon Technologies, Illumina, and Oligos, etc., for primer synthesis. Primer is delivered at room temperature in pellet form. We order at the 50 nmol scale with no additional purification.
9. Epicentre technologies does not provide specific information on reagent concentrations for the Failsafe 2X premixtures. The "midrange" refers to premixtures "D," "E," "F," and "G."
10. Often a PCR thermal cycler program utilizing a two-step or touchdown method is superior. This methodology allows increased specificity of primer annealing by carrying out the first 10 cycles at a fairly high annealing temperature (e.g., 60°–65°C) and the remaining cycles at a temperature approx 7°–10°C lower, aimed at boosting the yield (e.g., 50°–58°C).
11. Optimization of PCR largely depends upon the gene under investigation, espe-

cially GC content and inherent diversity. Ultimately, we have found that the more difficult a particular inbred line is to amplify, the more interesting the nucleotide sequence. Quite often, certain inbred lines require separate optimization to obtain PCR product, and even then, sequencing can require "line-specific" primers where highly polymorphic regions exist. We strive to obtain 2× coverage over the entire gene for almost all the samples.

12. We gauge the bandwidth to determine appropriate elution vol on purification by rough visual quantification. Mainly, we check to ensure there is a single band of the expected size present. Typically, we use 30 ng DNA/reaction for sequencing. Even less concentrated product may yield adequate sequence results. Only very weak bands, as visualized by gel electrophoresis, will yield poor results (e.g., <15 ng DNA/8 μL PCR product).
13. PHRED and PHRAP allow base calling and assembly of DNA sequence by simple Fourier methods.
14. Gaps are treated as missing data, and all sites at those positions are excluded from analyses; gaps in exons may alter the translation.
15. For example, in maize we use *Tripsacum*, which diverged from maize about 5 million yr ago, while the allelic diversity in maize is roughly 1 to 2 million yr old.
16. If inbred lines are being used, we set the second allele to missing (–9) as it eliminates the Hardy-Weinberg part of the model, and helps reconstruct the population structure before the inbreeding.
17. Sometimes the model seems to split off individual taxa, however, these single taxa populations are not very useful for controlling population structure. The user may want to try the Q matrix based on the population number before the individual taxa populations are split off.
18. The described approach analyzes individual sites, however, the analysis of haplotypes can also be powerful statistically. There are many approaches in the human genetic literature that could be used.
19. The SAS script for the test is as follows:

```
proc logistic data = indata outest = resultH0;
model testPolymorphism = Q1 Q2; run;
proc logistic data = indata outest = resultH1;
model testPolymorphism = trait Q1 Q2; run;
```

Then the difference of _LNLIKE_ of both tests is used as the test statistic.

Acknowledgments

We thank Jeffry Thornsberry, Brad Rauh, Sandra Andaluz, Sherry Flint-Garcia, Susan Wiltse, and Larissa Wilson for helping to develop these methods and commenting on this manuscript. This research was supported by National Science Foundation (NSF) grant DBI-9872631 and the United States Department of Agriculture–Agricultural Research Service (USDA-ARS).

References

1. Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S., IV. (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* **28,** 286–289.
2. Remington, D. L., Thornsberry, J. M., Matsuoka, Y., et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* **98,** 11479–11484.
3. Nordborg, M., Borevitz, J. O., Bergelson, J., et al. (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30,** 190–193.
4. Sharbel, T. F., Haubold, B., and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9,** 2109–2118.
5. Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65,** 220–228.
6. Reich, D. E. and Goldstein, D. B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* **20,** 4–16.
7. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959.
8. Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* **67,** 170–181.
9. Hudson, R. R., Kreitman, M., and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159.
10. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185.
11. Rozas, J. and Rozas, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15,** 174–175.
12. Hey, J. and Wakeley, J. (1997) A coalescent estimator of the population recombination rate. *Genetics* **145,** 833–846.
13. Schneider, S., Roessli, D., and Excoffier, L. (2000) *Arlequin ver. 2.000: A Software for Population Genetics Data Analysis*. Genetics and Biometry Laboratory, University of Geneva, Switzerland.
14. Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics and breeding. *Curr. Opin. Plant Biol.* **5,** 94–100.
15. Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38,** 226–231.
16. Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA* **98,** 9161–9166.
17. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.
18. Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–971.