

# Chapter 10

## Exploiting Barley Genetic Resources for Genome Wide Association Scans (GWAS)

Robbie Waugh, Andrew J. Flavell, Joanne Russell, William (Bill) Thomas, Luke Ramsay and Jordi Comadran

### Contents

10.1	Introduction . . . . .	238
10.2	Multi Parent Populations . . . . .	239
10.3	Linkage Disequilibrium . . . . .	239
10.4	Population Structure . . . . .	241
10.5	Genetic Markers . . . . .	244
10.6	Ascertainment Bias . . . . .	245
10.7	GWAS . . . . .	247
10.8	Future Prospects . . . . .	250
	References . . . . .	251

**Abstract** We have been exploring the use of GWAS for trait analysis and gene isolation in cultivated barley. In this chapter we describe the approach we have taken and some of the hurdles that we have faced when attempting to establish the whole system. We discuss the way that we, but also others, have addressed the various issues that have arisen and provide guidance on how they can be avoided. These range from choosing the appropriate population for analysis, how to deal with inherent population structure, genetic marker discovery, application and the effect of ascertainment bias to the range of software currently available for conducting association analyses. We conclude by providing a series of successful examples from our laboratory that range from analysis of simple single gene traits through oligogenic to quantitative traits, and the detection of epistatic interactions. We conclude that appropriately designed and executed GWAS in barley is a powerful tool in our quest to identify the genes and alleles underlying key genetic traits.

---

R. Waugh (✉) · J. Russell · W. (Bill) Thomas · L. Ramsay · J. Comadran  
The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland  
e-mail: Robbie.Waugh@hutton.ac.uk

A. J. Flavell  
Division of Plant Sciences, The University of Dundee at JHI,  
Invergowrie, Dundee, DD2 5DA, Scotland

**Keywords** Barley · Germplasm · Linkage disequilibrium · Association mapping

## 10.1 Introduction

Crop plants evolved from their wild ancestors by the processes of domestication and selective breeding over the last *ca.* 10,000 years. Initially, wild plants carrying promising traits were cultivated, leading eventually to locally adapted landraces. These lost many undesirable alleles as useful alleles became enriched (Feuillet et al. 2008). Modern breeding has largely extended this by a process of crossing the ‘best with the best’ and the successes have been impressive. Unfortunately, there are indications that we are approaching a performance ceiling for at least some crops, as the best alleles become assembled in elite genetic materials (Tanksley and McCouch 1997, <http://www.fao.org/ag/agp/agpc/doc/riceinfo/Asia/ASIABODY.HTM>). The potential to re-invigorate these elite materials may be provided by the introduction of new alleles from wild species and old, locally adapted germplasm. Many studies have demonstrated the value of alleles originating from un-adapted and unimproved germplasm showing that centuries of selective breeding have not necessarily resulted in the accumulation of all the optimal alleles. For example, several barley cultivars have been released in Europe that contain fungal resistance genes introgressed recently from *H. spontaneum* (von Korff et al. 2005; Schmalenbach et al. 2009). A major challenge for the future is to streamline this process using high throughput genomics approaches.

The identification and recruitment of useful alleles are two very different tasks and both are difficult. Allele identification requires detailed and careful phenotypic trait analysis, combined with high-resolution genomic characterisation. Comparison between the phenotypic and genotypic data sets, either by linkage mapping in biparental populations or by genome wide association scanning (GWAS) of panels of related genotypes can in principle yield candidate marker alleles linked to the traits investigated. While the former approach has been generally successful in identification, deployment of the results in breeding has not been as widespread for many reasons, including the problems in identifying markers sufficiently closely linked for effective use in selection. The latter approach is therefore becoming more attractive because it is intrinsically higher resolution and, has the potential at least, to be more powerful because it scrutinises the results of many more generations of recombination and selection (Caldwell et al. 2006; Rostoks et al. 2006; Cockram et al. 2010). However there are also issues with GWAS that need to be resolved before it can be most effectively applied. In this chapter we will review some of the challenges that we have encountered and that need to be considered when planning to exploit genetic resources for GWAS. These are largely based on our experiences in establishing a successful GWAS programme in barley.

## 10.2 Multi Parent Populations

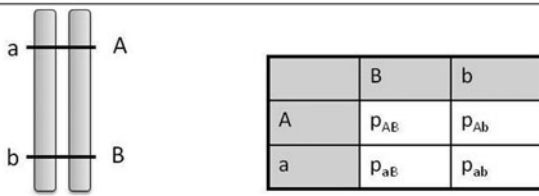
Over the past 25 years the correlation of phenotypic data with genetic markers in the offspring from specific bi-parental crosses using the well-established methods of ‘genetic linkage analysis’ has significantly advanced our understanding of the number, organization, location, and contribution of genetic loci to both simple and complex phenotypes (e.g. Turner et al. 2005; Yan et al. 2006). In a growing number of cases, particularly for Mendelian (i.e. single gene) traits, linkage mapping in very large populations has allowed the responsible genes to be fine mapped and ultimately to be cloned and analysed at the sequence level (Komatsuda et al. 2007). This has been achievable because the large number of recombination events in such populations allows the trait gene to be positioned so accurately that it is often possible to resolve its location to a specific DNA sequence (when available) or a single large-insert DNA clone that contains only one or perhaps a few candidate genes. Successes include major disease resistance and developmental genes such as *Mlo*, *Rpg1*, *Vrn1* and *Ppd1* and more will continue into the future. Bi-parental mapping requires the construction of specific populations that segregate for the trait of interest and because it samples only a small portion of the genetic variation inherent in the genepool under study, different populations are frequently required for each new trait studied.

More recently, geneticists have started to investigate GWAS in an attempt to increase the resolution of primary genetic studies. In contrast to linkage analysis, association approaches evaluate the correlation between loci and/or markers in populations of plants that share a degree of common history. Populations used for GWAS include collections of related individuals within natural or constructed populations from within a species. Association mapping effectively increases the number of recombination events to include all occurrences within the history of the sample. This presents a distinct advantage over bi-parental populations by improving genetic resolution from the megabase to the kilobase scale. The resolution inherent in a population used for GWAS is largely dependent upon the phenomenon of linkage disequilibrium a measure that can itself be complicated by the history of the population and which has the potential to increase the frequency false positive associations.

## 10.3 Linkage Disequilibrium

Linkage Disequilibrium (LD) is defined as the non-independence of alleles at different loci in a population (Box 1). At its most basic level, LD is maintained as a balance between mutation and recombination. At the moment of spontaneous (or induced) generation all new mutations are in perfect association with their genetic background. However, over time the processes of recombination (during meiosis) and genetic drift gradually lead to decay in the extent of these original associations and as new mutations are generated and selected, and old ones are lost, new associations are established. LD is therefore the product of evolutionary and biological factors that together contribute to the genetic structure and allelic histories of each

**Box 1**



For 2 loci each with two alleles, A and a at the first locus and B and b at the second, LD between these loci is given by:

$$R^2 = (p_{AB}p_{ab} - p_{Ab}p_{aB})^2 / p_A p_a p_B p_b$$

Where:

$p_A$  etc is the frequency of allele A in the population

$p_{AB}$  is the frequency of individuals with A allele at first locus and B allele at the second locus

**Note:**  
 $R^2$  measures statistical association and there is a simple inverse relationship between this measure and the sample size.  
 $R^2$  takes a value of 1 if only two haplotypes are present.

gene in the population. The extent of LD can be measured effectively by assaying and correlating the allelic state of genetically linked molecular markers at known genetic loci across the genome in what has been termed an association mapping panel of genotypes. When LD is extensive, statistically significant associations (correlations) may be detected between markers that are several to many centi-Morgans (i.e. potentially several megabases) apart. When it is low, associations between genes or markers may rapidly reduce to become non-significant at the sub-centiMorgan scale, or over thousands or even hundreds of bases. Within this generalised assertion, false positive associations can arise from the effects of genetic structure in the population, which may have originated from non-random mating, population bottlenecks or directional selection. As an example, up to 80% of the significant associations detected between polymorphisms in the maize *dwarf8* (*d8*) gene and flowering time were assessed as being due to population substructure (Thornsberry et al. 2001).

Mating system has a similarly profound impact on LD. Simulation studies have demonstrated that in the absence of mitigating factors, high levels of LD persist to a greater extent in highly selfing species (like barley), and that this is predominantly a factor of the *effective* recombination rate. This is simply because inbreeding results in increased homozygosity. Subsequently, as a consequence of this high homozygosity, a significant proportion of all recombination events in an inbreeding species will fail to bring about an exchange of genetic variation. This has been countered to some extent by artificial outcrossing, the basis of plant breeding practiced over the last hundred years. Therefore, in inbreeding crops like barley, while we would naturally

expect LD to be extensive in natural populations, plant breeding has been effective at generating a pseudo-outcrossing population where LD has been reduced to an extent that makes it useful for medium resolution association-based approaches and the identification of correlations between trait genes and alleles at molecular marker loci (Rostoks et al. 2006).

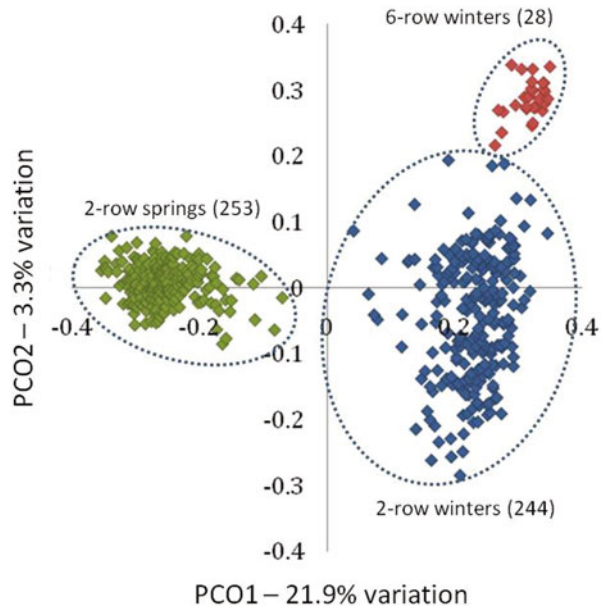
While it is relatively easy to detect marker-trait associations if there is extensive LD this inevitably results in a lower resolution map that requires more work to pin down the allele associated with the trait under study. Natural populations (including both true wild plants and adapted cultivated landraces) contain high levels of genetic diversity and are a great potential reservoir of DNA variation for crop improvement. Because of their history (i.e. number of generations), they also exhibit less extensive LD (Morrell et al. 2005; Kraakman 2005; Caldwell et al. 2006). These are potentially valuable as populations with low LD provide an opportunity to reveal high-resolution associations. Of course, if a genome wide approach is being adopted, the number of markers needed to find any associations would need to be extremely high, which is an associated cost. This has led to the suggestion that, at least in principle, associations could be mapped to an approximate genomic location in germplasm where LD is extensive, then exact genomic regions could be saturated using progressively wider germplasm with correspondingly lower LD but higher marker densities around the established location of the causal gene. In practice this has not yet been achieved.

## 10.4 Population Structure

For association mapping, the underlying population structure can be a strong confounding factor, especially for traits that have driven the geographical or environmental adaptation of the germplasm set. From a practical point of view, considerable care therefore has to be taken in choosing germplasm, avoiding—if possible—the inclusion of strong population stratification given it is a source of false positive associations. In other words, for a specific trait if there were major loci associated with genetically distinct homogeneous clusters of lines, many background markers carrying alleles exclusive to the specific clusters are also going to be associated with the trait, even though they are not causal. Not surprisingly, a number of approaches have been used to minimise these effects.

**Statistical Approaches** Our genome-wide association mapping studies in barley (*Hordeum vulgare*) have forced us to confront the problem of population structure as a confounding factor. Barley germplasm is strongly stratified reflecting crop type (in terms of growth habit and spike morphology) and geographical origin, which is heavily linked to local adaptation of the germplasm (Fig. 10.1). For most studies, genotyping and phenotyping are conducted simultaneously. Thus, the exploration and statistical adjustment for stratification is generally conducted within the running time of a project and there is little scope for choosing a different set of lines if structure turns out to be a considerable problem. Moreover, after expensive and time-consuming data collection, a natural tendency is to want include as many data points

**Fig. 10.1** Population structure in the cultivated elite barley gene pool (523 lines with 890 non position-redundant SNPs). Three main clusters are evident based on the major biological divisions within the species



as possible in an analysis. Thus statistical approaches that correct and/or account for the effects of population structure within association scans have guided most of the research on GWAS for the last few years. Several different approaches have been proposed in the literature (Mackay and Powell 2007). Issues however arise when the number and identity of markers that remain significant after employing different statistical population structure correction methods are either inconsistent or remove known biological factors correlated at some level with the population stratification. This can result in uncertainty over what QTL to prioritise for further studies or to use as diagnostics in Marker Assisted Selection (MAS).

It is worth mentioning that in an association panel the ancestral marker allele frequencies are not known. Therefore even with saturated genome coverage, is it not possible to build a genetic map *de novo* using LD and then to use this as a framework for visualizing the location of QTL. Thus, a prior genetic map using one or several bi-parental populations needs to be built in parallel to the association mapping panel to estimate the genetic, or better physical, order of the markers in the genome, unless of course the genome sequence of the target species has been assembled. Some of the main approaches for dealing with structure are:

**Structured Association** Structured association uses multiple polymorphisms assayed throughout the genome to compute statistics that capture the underlying population structure of the germplasm—introducing non-independence between genotypes as a result of common genetic background. Statistics can be then modelled within a Mixed Linear Model (MLM) framework to account for multiple levels of relatedness due to historical population structure and kinship (Yu et al. 2006). Different

software/ statistical packages—for example R v 2.9.0 (<http://www.R-project.org/>), TASSEL v.3.0 (<http://www.maizegenetics.net>) or Genstat 14 (VSN International 2011)—provide different ways of correcting for population structure which can be used to assess which best suits your data. A variance covariance matrix containing coefficients of co-ancestry (kinship matrix) can be included in the mixed model to account for genetic relatedness between genotypes. Eigenanalysis uses the scores of the most significant principal components from the molecular marker matrix as co-variables in the mixed model, which is an approximation to the use of a kinship matrix. In barley, we found a mixed linear regression model (Yu et al. 2006), which accounts for multiple levels of relatedness due to historical population substructure and kinship, to perform best either implemented on its own and in combination with other methodologies. The significance threshold is usually estimated for each analysis using a Bonferroni corrected  $p$ -value of 0.05.

With the rapid increase of the amount of SNP marker data there is a need for methods that are able to cope with thousands to millions of computationally intensive analyses. To deal with this, emerging methodologies provide us with a choice of both approximate [e.g. GRAMMAR (Aulchenko et al. 2007), implemented in GenABEL (<http://www.genabel.org/packages/GenABEL>), P3D (Zhang et al. 2010), implemented in TASSEL (<http://www.maizegenetics.net/tassel>), EMMAX (Kang et al. 2010) (<http://genetics.cs.ucla.edu/emmax/>)] and exact methods [e.g. FMM (W. Astle & D. Balding, <http://www.genabel.org/MixABEL/FastMixedModel.html>), FaST-LMM (Lippert et al. 2011) (<http://mscompbio.codeplex.com/>), GEMMA [M. Stephens lab (<http://stephenslab.uchicago.edu/software.html>)] to account for structure effects.

**Naive Approach** In its simplest form, the *naive* approach—which does not account for any population structure correction—is based on the same principles that work for bi-parental QTL mapping populations and consists of a regression of the phenotype upon the genotype to detect the QTLs. Each marker in a genetic map has a probability to be associated with the QTL of interest. The naive approach is suitable for use in the following two types of population—though some would argue that as all populations have some residual structure, a structure correction should always be applied.

*Constructed Populations* New population types that capture the advantages of both linkage mapping and GWAS, and that focus on achieving high statistical power, high resolution and low population stratification have been developed in several species and have, or are, being developed in barley. Nested Association Mapping (NAM) (McMullen et al. 2009) and heterogenic stock inbred lines, also known as multi-parent advanced generation intercross or MAGIC populations overcome the handicaps imposed by stratification in natural germplasm collections (Cavanagh et al. 2008). Trait mapping using NAM and MAGIC populations is more complete due to greater genetic diversity and more precise than classical bi-parental populations. The short history of recombination gives high statistical power to QTL detection, while ancestral recombination and diversity accumulated between the parental lines provide the basis for much finer scale mapping. Rounds of inter-crossing and selfing remove long range LD present between the parental lines, and each extra generation

will shuffle the genetic contribution from the founder lines more and more. For NAM in Maize, twenty-five diverse lines were crossed to B73 and the F1 plants self-fertilized for six generations to create a series of twenty-five recombinant inbred line (RIL) families ultimately totalling 5000 individuals. In MAGIC populations a complex and time-consuming crossing scheme has to be implemented to avoid the creation of clusters of highly related progenies that could potentially introduce *de novo* germplasm stratification.

*Sub-Populations* Artificial out-crossing imposed by breeders coupled with the long recombination history of crop germplasm can create a highly diverse germplasm stock without major population sub-divisions. Assembling a population of this type is the approach we have taken. By exploiting the European elite two-rowed spring barley genepool, our association mapping population effectively behaves like a heterogenic stock inbred line population without strong stratification. It lacks confounding population effects and its assembly avoided complex and time-consuming crossing schemes. Most important from our point of view was that it enabled us to perform QTL analysis and discovery in a germplasm set that was directly related to the contemporary barley breeding genepool. We explored population structure in a large set of germplasm then used phylogeny, principle coordinates and STRUCTURE analyses to explore stratification and admixture in the germplasm, then chose to remove outlying lines from the final panel that we now use routinely for association mapping studies.

## 10.5 Genetic Markers

Given the increased resolution in association mapping panels to maximise the chances of exploiting it effectively, it is important that the number of molecular markers used for analysis is sufficient to exploit the number of recombination events. An early attempt at an association analysis in barley was by Kraakman and colleagues (2004). Using sparse genome coverage they reported a number of significant associations for yield and stability of yield with a number of AFLP loci. They claimed some correspondence of the position of these loci with known QTL from biparental mapping studies but this assertion was complicated by a lack of common markers. In a subsequent study using the same material they reported marker loci significantly associated with Barley Yellow Dwarf Virus resistance and quantitative measures of leaf rust resistance (Kraakman et al. 2006). Again some correspondence of positions with previous studies was claimed but in one instance the particular AFLP locus had been previously reported to be the peak marker for Rphq2, a major QTL for partial resistance to *P. hordei*. The most important limitation in these early studies was that the marker technology employed, AFLP, is not well suited to this application.

A breakthrough came with the development of highly parallel SNP assay systems such as the Illumina GoldenGate™ assay implemented with their oligo pool array technology (Fan et al. 2003; Rostoks et al. (2006) and Close et al. (2009)) used alignments between barley EST sequences to identify SNPs and used these to generate



two 1536 SNP barley oligo pool assays (BOPA1 and BOPA2). Using BOPA1 on a relatively small population of barley cultivars Rostoks et al. (2006) successfully identified associations between a cluster of CBF genes responsible for winter hardiness in barley by GWAS after classifying the genotypes according to their spring or winter growth habit. Since then, more dense arrays of markers have been produced for application in GWAS. For example, we recently exploited Illumina GAIIX RNA-seq datasets from a range of barley cultivars to identify > 30,000 robust SNPs and incorporated approximately 8,000 of these on a higher density SNP platform called a 9K iSELECT Infinium array (our unpublished results). It is likely that similar but higher density chips with > 30,000 SNPs will be developed in the near future.

However there is some debate over whether this platform is the best in the longer term. As the cost of generating high coverage genome sequence continues to drop, we and others have turned to another approach termed Genotyping-by-Sequencing (GbS) (Elshire et al. 2011). GbS promises even deeper depth of coverage of polymorphic sequence information while avoiding the serious issue of ascertainment bias inherent in SNP chip platforms (see below). The disadvantage at the current moment in time is that the informatics pipelines required to analyse GbS datasets require custom scripts, generally written by specialists in the labs pioneering the approach. In contrast, Infinium array development is accompanied by an 'out-of-the-box' software suite from the vendor that enables simple allele calling and QC along with easy export into various analytical packages. Of course, this situation will rapidly change as more individuals adopt the GbS approach.

## 10.6 Ascertainment Bias

The development of multiplex assays such as the Infinium chip discussed above generally involves mining data extracted from a limited number of individuals. The utility of the SNP sets thus obtained is affected by the parameters of this discovery protocol. SNPs are generally identified in a discovery panel, which consists of a small sample of individuals from a population. As this panel represents only a subset of the individuals, only a fraction of total polymorphisms will be discovered. Consequently, when these SNPs are then genotyped on a larger sample of individuals an 'ascertainment bias' is introduced (Nielsen 2000). Because the discovery panel is small, the probability that a SNP will be identified in this panel is a function of the allele frequency. Thus, rare SNPs will go undiscovered more often than common SNPs. When a SNP platform developed this way is then used to screen a much broader set of germplasm, the introduced bias may compromise measures of relatedness and genetic diversity. This is largely because statistical measures that rely on allele frequency, such as nucleotide diversity, population genetics parameters and linkage disequilibrium will be affected, and have been observed (Nielsen 2000; Schlotterer and Harr 2002; Rosenblum and Novembre 2007; Storz and Kelly 2008). In barley BOPA1, BOPA2 and the recent 9K iSelect platform have also been selected from a limited number of barley accessions (Rostoks et al. 2005, 2006; Close et al.

2009; Waugh et al. unpublished data). These SNPs have provided extensive genome coverage and have dramatically progressed our understanding of the distribution of genetic diversity within the barley gene pool. Indeed several large scale projects have already used these platforms to identify marker-trait associations in elite cultivars (AGOUEB, <http://www.agoueb.org>; BarleyCAP, <http://barleycap.cfans.umn.edu>; ExBarDiv: [http://pgrc.ipk-gatersleben.de/barleyNet/projects\\_exbarDiv.php](http://pgrc.ipk-gatersleben.de/barleyNet/projects_exbarDiv.php)) (Waugh et al. 2010). We should be mindful that the extent and patterns of diversity observed will be limited by such ascertainment issues present in the underlying data.

Particularly problematic is the use of SNPs ascertained from the cultivated gene pool to examine diversity outside of that genetically narrow set. In barley we are fortunate to have extensive collections of wild progenitors collected from the Mediterranean basin through south western Asia and eastwards as far as Tajikistan and the Himalayas, as well as locally cultivated landraces grown throughout the marginal regions of the Fertile Crescent. Understanding the genetic diversity within these, particularly the landrace collections that grow and yield under extreme conditions of temperature and water availability, will be important in future breeding programmes that seek to respond to a range of environmental challenges.

Moragues et al. (2010) evaluated the effects of SNP number and selection strategy on estimates of germplasm diversity and population structure for different types of barley collections. Using the 1536 BOPA1 SNP data and various subsets of 384 and 96 SNPs that could in principle be used for affordable middle-throughput genotyping platforms, they compared diversity statistics for 161 landraces from Jordan and Syria with 171 European cultivars. Differences were observed in patterns of SNP polymorphisms as well as a lower estimate of diversity in the landraces, contradicting previous studies using SSRs (Russell et al. 2003). This bias could be at least partially nullified by selecting an appropriate subset of SNPs. All marker subsets gave qualitatively similar estimates of the population structure in both landraces and cultivars. Russell et al. (2011) described the first application of the BOPA1 SNP platform to assess the evolution of barley in a portion of the Fertile Crescent, by genotyping geographically matched landrace and wild barleys (448 accessions) from Jordan and Syria. The question of ascertainment bias skewing the landrace-wild comparison, through greater 'pruning' of rarely polymorphic markers in wild germplasm and generating an underestimate of genetic diversity, was addressed. While they were unable to exclude this possibility, their data did show higher levels of genetic variation in wild material suggesting that the relative pruning of SNPs in wild compared to landrace barley is most likely limited. Furthermore, the difference in diversity levels between landrace and wild barleys was similar to that found in previous work (Russell et al. 2004).

In this particular study they wanted to examine diversity across the genome and particularly in regions that have been identified as playing a role in domestication. If the effect of bias, introduced by choosing SNPs polymorphic in elite cultivars was likely to be problematic, the result would be a reduction of diversity in wild compared to landraces around the domestication genes; countering the objective of the study. They identified 141 cases where rolling diversity estimates were significantly different between wild and landraces, with diversity higher in wild material the vast

majority (132 cases). Many were in regions of the genome where domestication genes are found. With the possibility of ascertainment bias pushing the comparison in the other direction, this result therefore becomes doubly significant.

## 10.7 GWAS

The feasibility of mapping Mendelian traits that are determined by single major genes by GWAS using panels of barley cultivars was clearly demonstrated by mapping SNP polymorphisms in germplasm collections by LD to positions that corresponded exactly to locations previously assigned by biparental genetic mapping (Rostoks et al. 2006; Waugh et al. 2010). This approach has been subsequently extended to analysis of simple and more complex phenotypic traits

**GWAS for Simple Phenotypes** In the first reported study, Kraakman et al. (2006) used a Pearson correlation coefficient between vectors of the phenotypic response and genetic markers, correcting for multiple testing and population structure, to identify a significant association between the DUS character ‘rachilla hair length’ and the microsatellite BMAG223. Subsequently, we used GWAS to investigate the morphological differences that are used for the characterisation of cultivars in tests of Distinctness, Uniformity and Stability (DUS). DUS characters form a ready source of highly heritable traits that are presumed to be under the control of a limited number of major genes. Cockram et al. (2010) used 490 cultivars (both winter and spring) that had been genotyped with BOPA1 revealing 1,111 sufficiently informative markers. GWAS using a mixed model to correct for population substructure identified fifteen traits that had clearly significant associations with specific genomic regions. The majority of these traits appeared to identify a single genetic locus. They included ‘seasonal growth habit’ (1H), ‘grain lateral nerve spiculation’ (2H), ‘grain aleurone colour’ (4H), ‘hairiness of leaf sheath’ (4H), ‘rachilla hair type’ (5H), ‘ear attitude’ (5H) and ‘grain ventral furrow hair’ (6H). The positions of several of these genetic positions coincided with the previously known locations for these morphological characters, others such as the 1H position shown for seasonal growth habit were unexpected. Of particular interest was a region on chromosome 2H that was found strongly associated with a number of anthocyanin based DUS characters. They noted that the Mendelian locus *ANTHOCYANINLESS 2* (*ANT2*) had been previously reported on chromosome 2HL based on studies involving biparental crosses. Similar mapping work, with a biparental population also genotyped with BOPA1 indicated that the map location of *ANT2* coincided with the position identified in the association panel. Then they derived a composite phenotype with two character states: absence of anthocyanin coloration in all recorded tissues (awns, auricles and lemma nerves), or presence in one or more of these structures. GWAS of the composite phenotype (absence of anthocyanin coloration in all recorded tissues or presence in one or more of these structures) found the genetic interval controlling this trait to lie between 93.5 and 103.7 cM on chromosome 2H, with the peak association ( $-\log_{10} p = 51.7$ , marker 11\_21175) at 96.8 cM.

Additional genetic markers were developed using co-linearity with rice chromosome 4 and Brachypodium (*B. distachyon*) chromosome 5, ultimately defining the ANT2 locus to within a 0.57 cM interval flanked the barley homologues of *LOC\_Os04g47110* and *LOC\_Os04g47020*. These flanking markers were used to identify a minimum tiling path of BACs across the interval that were then sequenced. The 260 kb interval contained eleven genes, of which eight were located at collinear positions in one or more related cereal genomes. Three gene models were identified between the flanking markers, including a strong candidate gene that showed high homology to genes at the *R/B* loci that encode proteins containing a bHLH DNA-binding domain, that have previously found to control anthocyanin pigmentation in maize.

Sequencing a 4.6 kb interval across the candidate gene *HvbHLH1* in a subset of 90 cultivars identified 69 polymorphisms arranged in 4 haplotypes, with haplotype 1 exclusive to 'white' varieties, while haplotypes 2-4 were associated with anthocyanin coloration in one or more tissues. The identified polymorphisms between the haplotype groups included eight synonymous and four non-synonymous variants, as well as a 16 bp deletion within exon 6 that results in truncation of the predicted protein upstream of the bHLH domain. Subsequent genotyping in the complete association panel established that the 16 bp deletion occurred in all cultivars lacking anthocyanin pigmentation, and not in cultivars in which anthocyanin is expressed in one or more tissues. Thus, GWAS for this Mendelian trait identified a region of the genome that with additional marker development could be reduced to only three genes, including a strong candidate gene that showed functional variation and was diagnostic for the trait (see Cockram et al. 2010 for further details).

**GWAS for Simple Traits Identifies Epistatic Interactions** Cockram et al. (2008) identified two epistatic loci controlling vernalisation requirement by GWAS. The panel consisted of 429 spring and winter barley varieties and was genotyped with S-SAPs and SSRs together with markers based on gene specific amplicons. The genetics of vernalization requirement in barley is relatively well characterized being controlled predominantly by two major loci: *VRN-H1* and *VRN-H2* (von Zitzewitz et al. 2005). Spring alleles are thought to be due to deletions spanning putative *cis*-elements in *VRN-H1* intron I, or to deletions of part or all of the genomic region carrying the *VRN-H2* candidate genes. There is thus an epistatic relationship between the loci with winter barleys requiring winter alleles at both *VRN-H1* and *VRN-H2* potentially making their detection problematic in GWAS. However markers for both loci were found associated with winter habit in this panel with the use of genomic control (Cockram et al. 2008) as well as allowing for population structure in the analysis. This finding confirmed the results of previous detailed bi-parental mapping studies that had furnished the GWAS investigation with the markers targeting the functional polymorphisms at *VRN-H1* and *VRN-H2*.

The lack of genomic marker coverage hampered the study of Cockram et al. (2008). Ramsay et al. (2011) used the BOPA1 and BOPA2 platforms to elucidate the control of another epistatic interaction that aligns with population sub-structure in barley; that underlying ear-row number. Barley possesses three single-flowered

spikelets at each rachis node with the alternating triplets appearing opposite each other in two ranks thus forming six files of spikelets. When all three are fertile the ear has six rows of grains but if the two outer lateral spikelets are sterile then the ear is two-rowed. The presence of six rows is controlled principally by the cloned gene *VRS1*, on chromosome 2H (Komatsuda et al. 2007) that has been known for some time to be modified by the action of *INT-C* on chromosome 4H. In germplasm surveys, the *vrs1.a* allele in six-rowed barley cultivars is generally complemented by the *Int-c.a* allele and in two-rowed cultivars *Vrs1.b* is always complemented by *int-c.b*. The presence of *int-c.b* in six-rowed cultivars (i.e. *vrs1.a*, *int-c.b*) results in the development of smaller lateral spikelets (Lundqvist et al. 1997). In normal two-rowed (i.e. *Vrs1.b*) barley, *int-c.b* suppresses anther development in the lateral spikelets. In contrast, *Int-c.a* in two-rowed cultivars (i.e. *Vrs1.b*, *Int-c.a*) causes enlarged, partially male fertile, lateral spikelets.

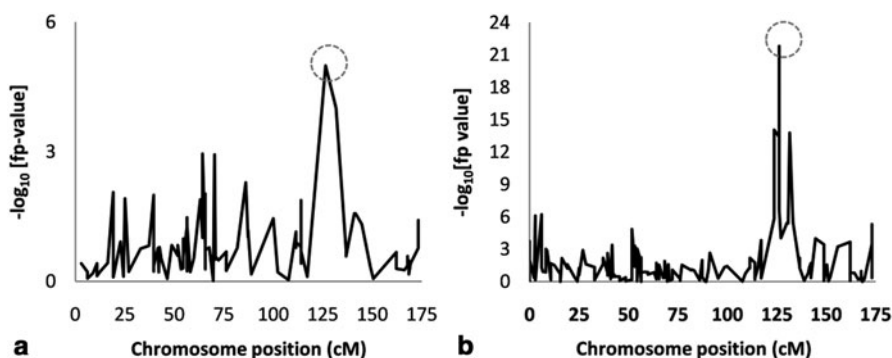
Row type is indicative of a major population division in barley germplasm, though some cross breeding has occurred, in particular in the development of European winter-sown barleys. Despite this population stratification, association tests of row type in 190 barley cultivars with 2473 bi-allelic genome-wide SNPs revealed associations on chromosomes 1HL, 2HL and 4HS. The association of a SNP in a gene estimated to be 0.05 cM (seven genes) distal to *VRS1* indicated that the peak on 2HL was caused by *VRS1*. This was confirmed by re-sequencing *VRS1* across the mapping panel, finding complete association with causal *vrs1.a* alleles. Direct evidence for the correspondence between the association on 4HS with *INT-C* was again complicated by a lack of common markers with previous mapping studies and the inherent difficulty in phenotyping the environmentally sensitive *intermedium* trait in bi-parental populations (Lundqvist et al. 1997). Using rice gene content and order as a proxy, further characterization of the region was once again achieved by re-sequencing PCR amplicons derived from barley orthologues of the neighboring rice genes across the association panel. This showed that a significant level of association was maintained over a region of some twenty genes that included several strong candidate genes for *INT-C*, notably the barley orthologue of maize *TEOSINTE BRANCHED 1* (*ZmTB1*). *ZmTB1* is a domestication gene and member of the TCP gene family that encodes putative basic helix-loop-helix DNA-binding proteins and whose members are involved in the control of organ growth. Resequencing confirmed that *HvTB1* contained the most significantly associated SNP and genetic mapping that placed it in the expected location. Definitive evidence that *HvTB1* was *INT-C* was obtained by re-sequencing *HvTB1* in a collection of 17 known *INT-C* mutants in a *Vrs1.b* (two-row) background. The GWAS approach thus enabled dissection of the epistatic control of row-type and high resolution mapping, and ultimately cloning of the interacting genes.

**GWAS for Quantitative Traits** The use of GWAS to dissect the genetic control of quantitative traits is more complex than its use for simpler traits controlled by a limited number of major genes. There are evident limitations to the power of a GWAS to determine the loci underlying a quantitative trait depending on the size and nature of the panel used as well as the complexity of the genetic control of the trait. Simulations can give some guidance to the expected limitations of the

power of a particular study (Cockram et al. 2010) as well as to the appropriateness of methodologies to allow for population structure. However, the use of a much higher density of markers and the direct relationships established between association and bi-parental studies revealed by sharing same genotyping platform have made such comparisons easier in recent studies. The functional validation of candidate genes underlying quantitative variation is more complicated than those under the control of monogenic or oligogenic traits where developmental or morphological consequences of functional genetic variation may already have been characterised through the use of mutant plant resources. Usually the knowledge of the genetic architecture of the trait in the germplasm under study is scarce, there is no reference in bi-parental populations and even when positional correspondence between bi- and multi-parent populations is observed, it is generally difficult to prove that they share the same underlying genetic determinants. The nature of the trait may hinder exploration and using rice or *Brachypodium* gene content and order as a proxy is difficult because the type of gene responsible for the trait is maybe unknown. The most robust associations for entering the validation pipeline can be prioritised by identification of the same associations in independent germplasm. Figure 10.2 shows how a significant height QTL on chromosome 3H detected in a spring barley association panel consisting of 650 lines with *de novo* height data is cross-validated in an independent dataset consisting of 230 spring lines using 15 years of historical data. The association on chromosome 3H is almost certainly due to the green revolution gene *sdw1* (Jia et al. 2009) and is co-located with the *sdw1* phenotype mapped in a mapping populations (Thomas et al. unpublished data; Malosetti et al. 2011), but other associations observed have not yet been characterised. Given the difficulties associated with validating associations with components of complex traits it is not surprising that there is little in the literature yet describing successes in this domain. However the authors are aware of several studies where components of complex traits have been resolved to gene level and validated using mutant resources (Jordi Comadran and colleagues—unpublished results).

## 10.8 Future Prospects

Over the past several years we, and others, have successfully assembled the molecular tools, tested various analytical approaches and ‘tuned’ our choice of biological resources to effectively take advantage of genome wide association scans. Ultimately we chose to focus on exploiting variation in the relatively narrow 2-row spring barley genepool to take advantage of the limited population substructure, to reduce the number of segregating alleles at each locus, to facilitate generation of an efficient unbiased genotyping platform and to focus on contemporary germplasm that is still exploited for breeding in the public and private sectors. This latter choice in particular has allowed us to interact effectively with those involved in crop improvement and allowed easy transfer of resources and technologies into a domain that has real impact on determining the varieties that are grown in farmers’ fields. These choices



**Fig. 10.2** Cross-validation of genome wide association (GWA) scans across independent germplasm sets genotyped with the same SNP platform. BOPA1 SNP loci with minimum allele frequencies > 10 % and missing data < 10 % were used for a GWAS using a kinship mixed model approach as implemented in Genstat v.14 (VSN International). TASSEL V3.0 was used to estimate the kinship matrix (K) from a subset of random markers covering the whole genome so that we did not over-estimate sub-population divergence. (a) Highly replicated height data collected from 200 elite 2 row spring cultivars over a period of ~20 years were analysed by GWAS. Several significant association peaks were detected but only chromosome 3H is shown.  $-\log_{10}$  [fp values] are plotted following chromosomal order and may not reflect genetic distances. (b) Chromosome 3H scan for “*de novo*” height data collected on 650 2 row spring cultivars in one season. The top SNP (highlighted in the graphs with a circle) is tightly linked to barley green revolution gene *sdw1* (Ramsay et al. unpublished data)

together have allowed the isolation of major genes and genes controlling more complex traits. In future a significant issue remains over how we most effectively validate associations with components of highly complex traits such as yield and quality, and in such cases how the data is best exploited by the end user community. Thus, while as academics we are focused on using the information for gene identification and validation, we are also actively exploring how the phenotypic and molecular marker data can be integrated into a practical crop improvement program. Currently we are focusing on ‘Genomic Selection’ (GS—Meuwissen et al. 2001). A general view is that GS holds much promise for crop improvement but precisely how it will be implemented remains to be established. We conclude that, if establishing GWAS in barley effectively delivers the dual outcomes of facilitating gene isolation and providing the molecular and phenotypic datasets to establish Genomic Selection, then what we have learned will have been valuable and worthwhile.

## References

- Aulchenko YS, de Koning D-J, de, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177:577–585
- Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *hordeum vulgare*. *Genet* 172:557–567

- Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* 11:215–221
- Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdogan S, Roose ML, Moscou MJ, Varshney R, Chao S, Szücs P, Sato K, Hayes PM, Matthews DE, Marshall DF, Muehlbauer GJ, Graner A, DeYoung J, Madishetty K, Fenton RD, Condamine P, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582
- Cockram J, White J, Leigh FJ, Lea VJ, Chiapparino E, Laurie DA, Mackay IJ, Powell W, O'Sullivan DM (2008) Association mapping of partitioning loci in barley. *BMC Genet* 9:16
- Cockram J, White J, Zuluaga D, Smith D, Comadran J, Macaulay M, Luo ZW, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Waugh R, O'Sullivan D (2010) Genome-wide association mapping of morphological traits to candidate gene resolution in the un-sequenced barley genome. *Proc Natl Acad Sci USA* 107:21611–21616
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell Sharon E (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species *PLOS ONE* 6:e19379
- Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelplnagel J, Chee MS (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* 68:69–78 (vol LXVIII)
- Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. *Trends Genet* 24:24–32
- Jia Q, Zhang J, Westcott S, Zhang XQ, Bellgard M, Lance R, Li C (2009) GA-20 oxidase as a candidate for the semidwarf gene *sdw1/denso* in barley. *Funct Integr Genomics* 9:255–262
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–U110
- Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genet* 168:435–446
- Kraakman ATW (2005) Mapping of yield, yield stability, yield adaptability and other traits in barley using linkage disequilibrium mapping and linkage analysis. PhD dissertation 3772. Wageningen University
- Kraakman ATW, Martí'nez F, Mussiraliyev B, van Eeuwijk FA, Niks RE (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. *Mol Breed* 17:41–58
- Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, Lundqvist U, Fujimura T, Matsuoka M, Matsumoto T, Yano M (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc Natl Acad Sci USA* 104:1424–1429
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* 8:833–U94
- Lundqvist U, Franckowiak JD, Konishi T (1997) New and revised descriptions of barley genes. *Barley Genet Newsl* 26:22–516
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Malosetti M, van Eeuwijk FA, Boer MP, Casas AM, Elia M, Moralejo M, Ramsay L, Molina-Cano JL (2011) Gene and QTL detection in a three-way barley cross under selection by a mixed model with kinship information using SNPs. *Theor Appl Genet* 122:1605–1616
- McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes K, Kroon D, Lepak



- N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas OM, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) "Genetic properties of the maize nested association mapping population". *Science* 325(737):737–740
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genet* 157:1819–1829
- Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high throughput SNP genotype data. *Theoretical And Applied Genetics* 120:1525–1534
- Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc Natl Acad Sci USA* 102:2442–2447
- Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–942
- Ramsay L, Comadran J, Druka A, Marshall DF, Thomas WTB, Macaulay M, MacKenzie K, Simpson CG, Fuller J, Bonar N, Hayes PM, Lundqvist U, Franckowiak JD, Close TJ, Muehlbauer G, Waugh R (2011) Intermedium-C, a modifier of lateral spikelet fertility in barley is an ortholog of the maize domestication gene *teosinte branched 1*. *Nature Genetic* 43:169–172
- Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *J Hered* 98:331–336
- Rostoks N, Mudie S, Cardle L, Russell JR, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Mol Genet Genomics* 274:515–527
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole genome association mapping in elite crop varieties. *Proc Natl Acad Sci USA* 103:18656–18661
- Russell JR, Booth A, Fuller JD, Baum M, Ceccarelli S, Grando S, Powell W (2003) Patterns of polymorphism detected in the chloroplast and nuclear genomes of barley landraces sampled from Syria and Jordan. *Theor Appl Genet* 107:413–421
- Russell J, Booth A, Fuller F, Harrower B, Hedley P, Machray G, Powell W (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the VSN International, Hemel Hempstead barley genome. *Genome* 47:389–398
- Russell JR, Dawson IK, Flavell AJ, Steffenson B, Weltzien E, Booth A, Ceccarelli S, Grando S, Waugh R (2011) Analysis of more than 1,000 SNPs in geographically-matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol* 191:564–578
- Schlotterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Molecular Ecology* 11:947–950
- Schmalenbach I, Léon J, Pillen K (2009) Identification and verification of QTLs for agronomic traits using wild barley introgression lines. *Theor Appl Genet* 118:483–497
- Storz JF, Kelly JK (2008) Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. *Genet* 180:367–379
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Sci* 277:1063–1066
- Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 28:286–289
- Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Sci* 310:1031–1034
- VSN International (2011) *GenStat for Windows*, 14th edn. VSN International, Hemel Hempstead, UK. Web page: [GenStat.co.uk](http://GenStat.co.uk)

- von Korff M, Wang H, Léon J, Pillen K (2005) AB-QTL analysis in spring barley. I. Detection of resistance genes against powdery mildew, leaf rust and scald introgressed from wild barley. *Theor Appl Genet* 111:583–590
- von Zitzewitz J, Szücs P, Dubcovsky J, Yan L, Francia E, Pecchioni N, Casas A, Chen THH, Hayes P, Skinner J (2005) Molecular and structural characterization of barley vernalization genes. *Plant Mol Biol* 59:449–467
- Waugh R, Marshall D, Thomas WTB, Comadran J, Russell JR, Close T, Stein N, Hayes P, Muehlbauer G, Cockram J, O'Sullivan D, Mackay I, Flavell AJ, Agoueb, BarleyCAP, Ramsay L (2010) Whole-genome association mapping in elite inbred crop varieties. *Genome* 53:967–972
- Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci USA* 103:19581–19586
- Yu J, Pressoir G, Briggs WH, Vroh I, Bi M, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordoñas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42:355–U118