

Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants



James Cockram and Ian Mackay

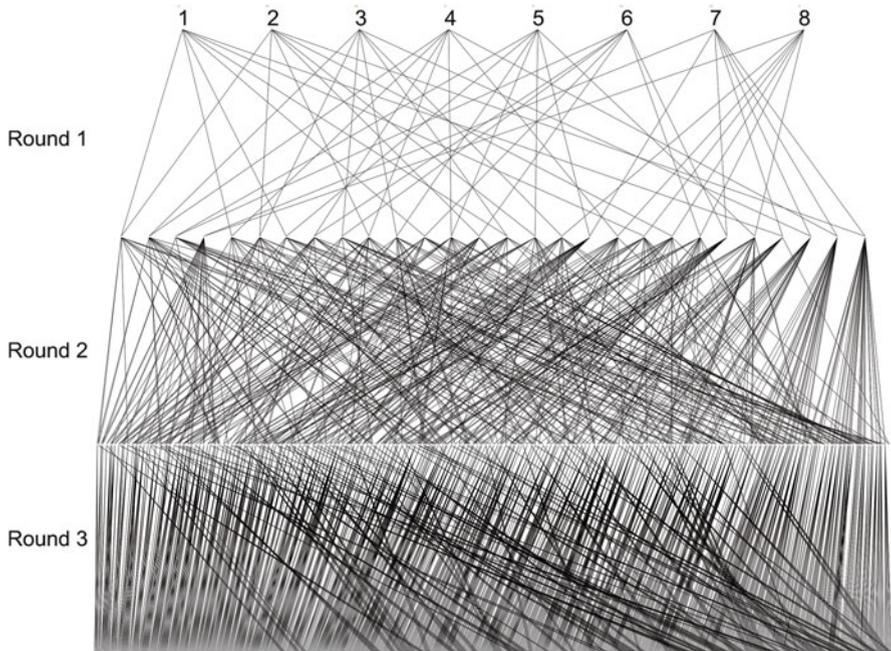
Abstract Fine mapping of quantitative trait loci (QTL) is the route to more detailed molecular characterization and functional studies of the relationship between polymorphism and trait variation. It is also of direct relevance to breeding since it makes QTL more easily integrated into marker-assisted breeding and into genomic selection. Fine mapping requires that marker-trait associations are tested in populations in which large numbers of recombinations have occurred. This can be achieved by increasing the size of mapping populations or by increasing the number of generations of crossing required to create the population. We review the factors affecting the precision and power of fine mapping experiments and describe some contemporary experimental approaches, focusing on the use of multi-parental or multi-founder populations such as the multi-parent advanced generation intercross (MAGIC) and nested association mapping (NAM). We favor approaches such as MAGIC since these focus explicitly on increasing the amount of recombination that occurs within the population. Whatever approaches are used, we believe the days of mapping QTL in small populations must come to an end. In our own work in MAGIC wheat populations, we started with a target of developing 1,000 lines per population: that number now looks to be on the low side.

J. Cockram (✉) and I. Mackay

The John Bingham Laboratory, National Institute of Agricultural Botany (NIAB), Cambridge, UK

e-mail: james.cockram@niab.com

Graphical Abstract



Keywords Arabidopsis multi-parent recombinant inbred line (AMPRIL), Fine-mapping, Genome wide association scans (GWAS), Linkage disequilibrium (LD), Multi-founder advanced generation inter cross (MAGIC), Nested association mapping (NAM), Power, Precision

Contents

1	Introduction	111
2	The Importance of Mapping Trait Loci	111
3	The Importance of Fine Mapping	112
4	Factors Determining Precision in Fine Mapping	113
4.1	Recombination	113
4.2	Population Size	114
4.3	Size of Effect	114
5	The Need for Replication	115
6	Association Mapping	115
7	Genome-Wide Association Studies (GWAS) Compared to Experimental Populations for Fine Mapping	116
8	Experimental Populations	118
9	Biparental Populations	119
10	Bulk Segregant Analysis (BSA)	120
11	Advanced Intercross (AIC)	121

12	Near Isogenic Lines (NILs)	122
13	Multi-Founder Populations	123
	13.1 Nested-Association Mapping	124
	13.2 Heterogeneous Stock	125
	13.3 MAGIC	126
14	Analysis Approaches for MAGIC	129
	14.1 Arabidopsis Multi-Parent Recombinant Inbred Line (AMPRIL)	131
	14.2 Linked or Multiple Mapping Populations	132
15	Conclusion and Outlook	132
	References	134

1 Introduction

In this chapter, we review the use of mapping populations for precision location of quantitative trait loci (QTL). We focus on experimental populations created explicitly for trait mapping, distinguishing these from collections of lines or individuals used in the association-mapping approaches described in Chapter 5. We start by discussing trait mapping and the need for fine mapping, next describe some general properties and requirements of approaches to fine mapping, after which we describe specific approaches. We finish by considering prospects for the future.

2 The Importance of Mapping Trait Loci

The tagging of QTL with genetic markers has a long history [1]. However, most progress has been made since the 1980s with the development of DNA markers initiating the era of genome scans in bi-parental crosses [2, 3]. This has continued to the present, as new, cheaper classes of genetic markers (e.g., [4]) and improved statistical methods and software (e.g., [5]) have become available. Markers tagging QTL can be used to ease introgression of novel variation from un-adapted germplasm into elite lines, for marker-assisted selection, and for stacking multiple sources of disease resistance [6]. However, trait mapping has also been tempered by cautious voices, alerting people to the risks of bias in estimating effects [7], of lack of precision of QTL location [8], and of relevance to plant breeders' germplasm. Nearly 10 years ago Bernardo [9] commented on the very poor transfer rate of results from trait-mapping experiments to breeding programs. New population and genomics resources and more widespread understanding of the requirement for statistical power are now focusing effort on studies that map more QTL to smaller intervals in breeder-relevant germplasm. Our hope is that this will result in better application of QTL within breeding programs and other studies in the future, rather than being an end in itself.

3 The Importance of Fine Mapping

Fine mapping is the process by which the location of a QTL is reduced from an initial interval of 20 cM or more to an interval of a few cM or less. Taken to its final conclusion, fine mapping can lead to the identification of the causative genomic lesion, which could be a single nucleotide polymorphism, SNP, or other variant. Fine mapping has merit in biological studies: clearly if the causative polymorphism is identified, then the door is opened to further detailed molecular characterization and functional studies of the relationship between polymorphism and trait, and the wider genetic network. However, reduction of a QTL linkage interval to a smaller region without exactly identifying the functional polymorphism or gene is also of value. Near isogenic lines (NILs) can be created containing the alternative states at the QTL locus and used for detailed phenotyping. NILs cannot discriminate between pleiotropy and close linkage but the smaller the interval, the lower the chance of misinterpretation. For example, fine mapping in rice showed a rice photoperiod sensitivity locus, previously known as *Heading date 3* (*Hd3*), could be genetically dissected into two closely linked loci *Hd3a* and *Hd3b*, allowing allele-specific effects at both loci to be discriminated in NILs [10]. Fine mapping is also of direct relevance to breeding since it makes QTL more easily integrated into marker-assisted selection programs. A QTL located to an interval of 20 cM requires that a 20 cM tract of chromosome is tagged during selection and backcrossing. This prevents recombination within the region and the potential creation of favorable haplotypes containing the QTL and other, unidentified loci. In addition, wide intervals reduce the potential to stack favorable QTL since even with a small number of QTL, it is quite likely that intervals will overlap.

Genomic selection [11, 12] is the process whereby statistical models are applied to large numbers of genetic markers to predict the breeding (or trait) values of candidates for selection which are not themselves phenotyped. Because selection and phenotyping are decoupled, very high intensities of selection are possible (because single individuals rather than cultivars are selected) and reductions in cycle time can be made (because the breeding cycle is from parent to progeny, without time taken for cultivar development). In the near future, genomic selection will become routine in plant breeding [13]. QTL, known or recently discovered, can be incorporated into genomic prediction equations in an optimum manner, but a QTL characterized only by loosely linked flanking markers eliminates large tracts of chromosome from inclusion in trait prediction algorithms with potential loss of precision.

Identification of functional polymorphisms also allows genome-editing technologies, such as clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 (Cas9) [14], to introduce favorable changes directly into plant genomes without any subsequent linkage drag. Methods to increase the number of functional polymorphisms identified could have a substantial influence on rates of genetic improvement [15]. Breeding programs have already been studied in which genome editing is used to this end [15].

4 Factors Determining Precision in Fine Mapping

The precision with which a QTL can be located depends on three related factors: the recombination fraction between it and the available genetic markers, its heritability, and the size of the mapping population. These factors are general across all types of mapping populations. Other factors can affect specific approaches to fine mapping and will be discussed in context as they arise.

4.1 Recombination

Trait mapping in plants exploits the correlation between genetic markers and QTL. A zero correlation is expected between unlinked markers, rising to an expected maximum for a marker co-located with the QTL. However, the observed maximum is very unlikely to be located precisely at the QTL position. For example, for a marker and QTL separated by a recombination fraction of 0.01, the probability of observing at least one recombination in 100 meioses is only 63%; there is a very high chance that none occur, in which case a marker located at the QTL and a marker located ~1 cM away cannot be distinguished. In practice, the discrimination is likely to be much worse. With 1,000 meioses, the chance of observing no recombinations with a recombination fraction of 0.01, is 0.99996. Populations for fine mapping therefore require high levels of recombination. This is achieved in two ways. Firstly, population size can be increased, and secondly, the mapping population can be created over multiple generations of crossing. The early generations of selfing to generate recombinant inbred lines, when the frequency of double heterozygotes is reasonably high, also provide some opportunity for recombination. For this reason, fully inbred lines give greater precision than doubled haploid (DH) lines.

There is a conflict between power and precision in fine mapping. Power to detect a QTL is increased in populations with reduced recombination. This is exploited in standard bi-parental populations, but also, for example, in the pre-QTL mapping era by methods such as use of anisoploid lines (which possess a chromosome number that is an odd multiple of the haploid number, e.g., triploid) and whole chromosome substitution lines (e.g., [16]) which allowed the location of major effects (or the cumulative effect of many small effects) to whole chromosomes or chromosome arms. To increase precision, the mapping population requires more recombination events, but this comes at the expense of power. As a simple example: suppose a genome-wide significance threshold of 0.001 is required to detect a QTL linked to a marker located at a distance of 10 cM in an F_2 , and that additional cycles of crossing were used to increase precision to 1 cM. Ignoring complications from interval mapping and other multi-marker approaches, this would require a tenfold increase in marker density and therefore a tenfold change in the genome-wide Bonferroni corrected p -value to 0.0001. To maintain power to detect such an effect, population

size would need to be increased (see below). Adopting a lower significance threshold is an alternative, and this may be possible if the study population is not being used for GWAS, but to fine map a QTL identified in other studies.

4.2 Population Size

Population size has two effects. Firstly, as described above, bigger populations capture more recombination and therefore offer greater precision. However, in addition, bigger populations have greater power to detect QTL. For example, in a mapping population in which QTL and markers are segregating at a frequency of 0.5, for example a DH population or population of recombinant inbred lines (RILs), the power to detect a QTL accounting for 10% of the phenotypic variation between lines at a significance threshold of $p \leq 0.001$, and assuming a perfect marker for the QTL, is 0.52 with a population size of 100. Increasing the population size to 200 increases the power to 0.92. Increasing population size, therefore, increases both the precision with which phenotypic effects of marker classes are estimated and also increases the power to detect QTL. It is for this reason that increasing population size is preferred over increasing replicate number for a fixed population size. This latter approach increases the power with which QTL are detected but has less effect on precision.

4.3 Size of Effect

Bigger effects are easier to detect and are more precisely located. In the extreme, a QTL may be so large that, in practice, it behaves as a completely penetrant major gene: essentially as a marker itself. At the other extreme, for a highly polygenic trait, the effect of any individual locus may be very small and there will be very little power to detect or locate QTL. This is true even if the heritability of the trait is very high: it is the heritability of the QTL effect itself that is the dominant determinant of power and precision. However, for traits of low heritability, replication can be used to increase the heritability of line means and thus increase power and precision, though as previously mentioned precision is better increased by increasing recombination. There is a counter view, however [17], that conventional QTL mapping is very successful in locating QTL precisely, though the examples given are what most workers would regard as large QTL effects.

5 The Need for Replication

Whatever the interval reported or the method used, there is a requirement for replication, or validation, of the observed effect prior to progression of the QTL into marker-assisted breeding or functional investigation. Such studies should preferably be in a completely different population to that in which the QTL were first identified: it is not appropriate, for example, to partition a collection of germplasm, or a mapping population, and use one portion for “discovery” and the other for “validation.” Such an approach amounts to cross validation, and may reduce the rate of false positives *within that study* or result in reduced bias in the estimation of genuine effects through elimination of the winner’s curse [18], but it is not an independent study. The winner’s curse is the phenomenon that evidence of a new effect, provided by significance testing, often gives an inflated estimate of the size of that effect. In linkage analysis, this is often referred to as the Beavis effect [7]. Replication in independent studies has become a standard for publication of association-mapping studies in some journals [19] and has been more widely advocated, e.g., [20, 21]. In wheat, for example, we have identified 26 QTL associated with yellow rust resistance in a large panel of 488 lines, of which 11 out of 13 of the strongest associations were replicated in independent bi-parental mapping populations created for that purpose. Out of nine hits in the original association-mapping panel that were statistically significant at a less stringent significant threshold ($p \leq 0.001$), only three were replicated. Breeders were substantially more confident of incorporating the replicated QTL into their marker-assisted selection programs. Distrust of results from single studies contributes to the relative lack of uptake of QTL into marker-assisted selection programs [9].

6 Association Mapping

Association mapping, also known as linkage disequilibrium mapping, detects and locates QTL based on the strength of the association (correlation) between genetic markers and the traits under study. It relies on the magnitude of linkage disequilibrium (effectively the correlation) between genetic markers and QTL declining rapidly with genetic distance. Detection of a strong correlation between a trait and a genetic marker is therefore taken as evidence that a QTL is in close proximity to the marker. Association mapping can, in principle, be applied to any population or collection of lines or individuals, and in general will give higher precision than found in bi-parental populations. The pattern of LD decay is remarkably similar in diverse populations, due to the mechanism by which LD decays over time: at a rate of ($one - the\ recombination\ fraction$) per generation. However, differences in scale will be seen, due to differences between populations and species in the forces that

create LD – mutation, selection, and drift – and the age of the population or collection of lines under study with reference to their shared genealogy.

However, virtually all populations or panels assembled for association mapping include some degree of population structure or subdivision arising from ancient and very recent differences in the shared ancestry of the lines in the collection. If these are not taken into account, very high frequencies of false-positive results can arise: false in the sense that the observed association, though genuine, has arisen from some other cause than the close linkage of marker and QTL. Fortunately, statistical methods, in particular the use of the mixed model, can robustly adjust for population structure effects. For example, in the *Triticeae*Genome association-mapping panel of European wheat [22], 11% of squared correlations between markers pairs ≥ 0.8 were among unlinked markers. However, with simulated traits, application of the mixed model gave good control of false positives but identified that a higher marker density was required to improve precision. Power calculations, and associated estimates of expected precision, should always be reported in association-mapping studies. Association-mapping studies have therefore become routine in plants. For major genes, accuracy can be to the gene level (e.g., [23]), though independent evidence of the functionality of the candidate should also be obtained, for example through transformation or reverse genetics studies. For QTL that do not account for most of the genetic variation, accuracy is lower, but is generally greater than seen from bi-parental mapping populations. However, association mapping is not a panacea and other approaches still have a role, and may be better under some circumstances.

7 Genome-Wide Association Studies (GWAS) Compared to Experimental Populations for Fine Mapping

Genome-wide association studies (GWAS) have become increasingly popular in plants: they are easy to set up, requiring only the collection of pre-existing lines or cultivars with no need for de-novo crossing and selfing [24]. They also often come with pre-existing phenotypic data [23, 25]. However, there are a few notes of caution to be made:

GWAS studies are often proposed because of their assumed increased precision for locating QTL. However, this is often not realized. A demonstration that linkage disequilibrium (LD) decays quickly is usually used as an indication of likely precision in mapping, but if QTL are to be discovered in a genome scan, very large population sizes are required for fine mapping. Often, the sizes used in published studies are ludicrously small and there is no accompanying estimate of power. Occasionally, both population size and marker numbers are risibly low; 46 SSRs and 30 accessions with accompanying claims of high power and precision is the worst example we have seen published in an otherwise reputable journal, in this case with no report of the rate of LD decay or of power. Low power to detect a

QTL must throw doubt on published statements about the frequently large number of QTL detected [26]. This problem is not overcome by establishing, as often reported, that putative hits lie in known QTL linkage regions. Given the long history of QTL mapping, it is difficult for a locus not to lie in any region of the genome chosen at random. Statements, therefore, that say some proportion of the detected marker trait associations are new and the rest replicate previous results require greater statistical support, which can be easily calculated or simulated, but this is seldom done in plants. An excellent example in *Drosophila* is given by Highfill et al. [27]. They identified five QTL for variation in lifespan in a multi-parent advanced generation inter-cross (MAGIC) population. Over 100 QTL for this trait had previously been identified and they established through simulation that the probability that *all* of five randomly located QTL would overlap with one or more of these 100 was 0.85.

The problem of low power may be insurmountable. Population sizes can be unredeemably small, especially in the public sector where the only germplasm available may be from collections of varieties released commercially by breeders. If the number available is too low, the only option may be to create more, in which case an experimental population such as MAGIC may be a better alternative.

The statistical control of population structure and kinship works well in controlling the false-positive (Type I error) rate, but this can be at the expense of power. If a QTL is highly associated with a major population subdivision (e.g. [28]), it may be difficult to detect, or in extreme cases, not be detected at all: adjustments for kinship and population structure will reduce the power to detect any association between a QTL and linked marker that is also correlated with those effects. This can be a major problem: an attempt to increase power by capturing lines grown over a greater geographical and temporal range inevitably also increases population structure within the dataset. For example, the *Triticeae* Genome panel [22] includes lines of British, German and French origin. German lines tend to be taller: height being controlled by the use of growth regulators rather than through semi-dwarfing genes, and French lines tend to be earlier flowering – to avoid summer drought stress. Therefore, the frequency of major flowering time and height-reducing loci differs between countries so power to detect these is also reduced. In this case, these loci are of such great effect that they were still detected, though it is quite possible that minor QTL affecting these traits were not.

In addition, within narrow temporal and geographical ranges, although the problems of population structure may be reduced, LD may decay quite slowly with genetic distance; a consequence of close kinship among all lines. As a result, the precision with which QTL are detected can be reduced.

Power to detect and locate QTL in association-mapping panels also depends on the allele frequency of the QTL. Very rare alleles will not be detected in genome scans, even if the function polymorphism itself is tested and the effect is large. The detection of rare variants is an acknowledged problem in human genetics [20]. In plant science, we have the alternative of making experimental crosses and populations to reduce this problem, though success depends on the identification of appropriate founders.

Many of the issues surrounding the use of association-mapping panels for fine mapping can be avoided if the stringency of statistical significance is reduced. This is justified if the purpose of the experiment is not a GWAS but rather to test a small number of candidate genes or polymorphisms, or to fine map a genomic region for QTL identified in a previous study. For example, a genome scan with 10,000 markers would require a Bonferroni-adjusted significance threshold for an experiment-wide p -value of 5% of 0.0005%. Testing 100 candidate polymorphisms would require 0.05%.

In spite of the problems, association mapping is a powerful tool for fine mapping and notable successes have been reported [21, 23]. There is still of course a role for other methods: in particular using bespoke experimental populations.

8 Experimental Populations

Experimental populations for trait mapping pre-date the use of association-mapping panels, though, in essence, the principles are identical: QTL are located by the strength of the association between markers and traits, exploiting LD. In most experimental populations, LD decays slowly with genetic distance: a consequence of the limited number of generations between the creating of extensive LD by crossing a small number of founder or parental lines, and the small number of generations from that event before mapping takes place.

Experimental populations circumvent many of the problems with the use of association-mapping panels, but introduce problems of their own. There is little or no effect of population structure within them, so the correct Type I error rate is usually achieved. They tend to have higher power to detect QTL than association-mapping panels – a consequence of the slower rate of decay of LD with genetic distance. Correlated with this, they can lack precision in locating QTL in comparison to similar size association-mapping panels. In addition, they generally lack genetic diversity compared to association-mapping panels.

Selection of parents or founders is of great importance in experimental populations. This will be discussed in the context of individual population types below. As general principles, however, we note that parents or founders can be selected for similarity of phenological traits, in the hope of reducing segregational variation in the population for those traits to ease the phenotyping of the traits of major interest. For example, in many cereal species, lodging and resistance to abiotic stresses such as drought are affected by the developmental phase of the plants. Reducing segregational variation in, for example, flowering time, may increase both the power and precision with which QTL for stress are identified – in essence by increasing the heritability of the QTL for stress *per se*. However, matching parents on phenology is absolutely no guarantee that the problem is eliminated: dispersion of alleles with variable effects on phenology among the parents will generate substantial, and commonly transgressive, segregational variation in the progeny. In addition, selection of parents in this manner may eliminate

variation at linked loci, particularly if, as seems inevitable, there are interactions between phenology and other traits. The opposing view is to adjust for phenological variation during analysis by inclusion of covariates, or through partitioning the population into subsets with matched phenology (“slicing and dicing”), and then carrying out QTL analysis only within subpopulations. “Slicing and dicing” is equivalent to inclusion of covariates in the analysis to account for subgroup membership: in essence it amounts to binning a quantitative phenological factor such as flowering time into covariates with values of 0 and 1, and is therefore crudely equivalent to simply including phenology traits directly as covariates. Our opinion is that the inclusion of covariates is the better approach and is more flexible, plus it is likely that this will be required with matched parents anyway. However, as far as we are aware, the approaches have not been compared empirically or by simulation. In our own work, we have found the presence of transgressive segregation to be useful in interpretation rather than a hindrance in phenotyping.

9 Biparental Populations

Mapping in the progeny or lines from a cross between two inbred lines remains the standard approach for genetic mapping in plants (Chapter 4). Mapping usually takes place among DH lines derived from the F_1 , or RILs derived from the F_2 or backcross generations. Mapping directly among F_2 individuals is rarely used, partly because the genotypes of individuals are often poorly assessed by their phenotype (due to low heritability) and partly because, in the absence of clonal propagation, the individuals cannot be maintained indefinitely for annual or biennial species. The strength of the standard bi-parental population is in its power: LD decays slowly within chromosomes and there is no expectation of LD between loci on different chromosomes (LD between chromosomes is eliminated by the non-random mating of the parents: only the cross is made, not the selfs. With random mating, substantial LD would be found between chromosomes). As ever, the higher power is associated with low precision: Kearsey and Farquhar [8] state that precision for a standard QTL is to a region of 10–30 cM. They also make the point that the addition of markers to bi-parental mapping populations beyond a density of about one per 15 cM has limited effect on precision. The same point was made by Darvasi and Soller [29], who state that no increase in precision is made once an inter-marker recombination fraction of 0.1–15 is achieved. However, in practice more markers are commonly used: although only modest numbers of evenly spaced markers are required, it generally takes a much higher density of markers in the same population to produce an accurate genetic map in the first place. Moreover, reduced marker costs, SNP chips, and cheap genotyping by sequencing (GbS) will make discussions of marker density of historical interest. However, Kearsey and Farquhar [8] and Darvasi and Soller [29] point that it is more meiosis that is required to increase precision remains valid: small mapping populations, even if they have adequate power to detect QTL, lack in precision for QTL of modest size.

Selection of parents to create bi-parental mapping populations is generally trait driven: they are often selected as contrasting extremes for the trait of interest. This is a strength, in so far as it increases the likelihood that the population is segregating for multiple QTL for the trait. It also increases the chance of capturing alleles that are rare in the population sampled (and so are unlikely to be detected in GWAS studies). However, it is also a weakness in that the favorable alleles mapped may already have been fixed in breeders' germplasm. The trait-focused nature of many bi-parental mapping populations also means they tend to have short-term use. Interest in mapping additional traits is better served by creating additional targeted crosses, though there are exceptions of populations that have been more widely used. For example, the wheat Avalon \times Cadenza mapping population has been used a lot to map many traits [30]. Nevertheless, if inbred parents are sampled at random from a population, at most only half the loci segregating in the population will be segregating in the cross – assuming two alleles at equal frequencies at all loci – and it could be a much lower proportion.

One method of reducing the cost of creating bi-parental mapping populations is to use Rapid Bulk Inbreeding (RABID) [31, 32]. Here, inbreeding takes place in bulk, so is cheap, with modest numbers of markers (~ 100) used after inbreeding to identify a set of individuals with minimal relationships as the mapping population. As bulk inbreeding is cheap, it is possible to create multiple populations speculatively, but only genotype them, and multiply selected lines for phenotyping as required. A similar use of markers to select the most unrelated set of individuals could also be applied to lines produced by single seed descent or DHs, also providing the opportunity, for example, to eliminate lines carrying un-recombined chromosomes. Such chromosomes increase power but clearly do nothing for precision! We are not aware of any research looking at this.

10 Bulk Segregant Analysis (BSA)

Bulk segregant analysis (BSA) has had a revival now that DNA and RNA resequencing is routine in most plant species (reviewed for crops by Zou et al. [33]). The principle is that contrasting extremes are selected from a population and genotyped in bulks. Markers that distinguish between the bulks are judged to be closely linked to causative polymorphism. Bulking of the extremes saves money on genotyping, but is not necessary: genotyping individuals rather than bulks is more accurate.

There are three general approaches in terms of the biological materials used: (1) F_1 -derived individuals (F_2 s, RILs, DHs), e.g., identification of metabolite QTL for flavonoid production in *Arabidopsis* [34], (2) EMS-induced mutations, such as targeting induced local lesions in genomes (TILLING) populations, and (3) pooling genetically diverse breeding lines, e.g., red versus green hypocotyl in sugar beet [35].

Although the underlying principle of BSA is very simple, the expected composition of the selected pools, taking into account population size, number of selected individuals, genotype frequencies in the population, and genetic and environmental effects, is surprisingly complex [36]. Mackay and Caligari [37] used the methods of Hill [36] to compare BSA in F_2 and backcross generations, concluding that the F_2 was to be preferred if screening with dominant markers. There is, perhaps, a need to study strategies and population types more extensively, particularly in the light of DNA resequencing for genotyping: it is telling that the Hill [36] paper has only been cited four times.

11 Advanced Intercross (AIC)

The first experimental design intended specifically to improve precision in mapping was the Advanced Intercross (AIC) of Darvasi and Soller [29]. Acknowledging that absence of precision came from the limited recombinations captured in most bi-parental populations, they proposed increasing this by intermating the F_2 for several generations prior to mapping. For example, they stated that eight additional cycles of random mating could reduce the confidence interval of a QTL from 20 to 3.7 cM. Recent examples of AIC in plants include genetic investigation of female control of non-random mating in *Arabidopsis* (four rounds of intercrossing, 490 RILs, [38]) and disease resistance in maize (four rounds of intercrossing, 302 RILs [39, 40]). AIC, however, is used surprisingly rarely for crops. One reason for this may be the time required to make the additional crosses – especially if the mapping population was established in the first place only to detect (and locate) QTL for a specific trait in a short-term project, as is often the case. Another reason is that whereas a DH-mapping population may be created reasonably economically from an F_1 , since multiple DH lines can be created from a single individual, to do the same with an advanced intercross, or just an F_2 , requires, ideally, that a single DH is produced from as many outcrossed individuals as possible. Using the technologies currently available in many crops, this is not practical. Note that although the AIC will increase precision, this is at the expense of power. Essentially, this is a multiple testing problem. The number of independent genomic intervals, or their effective equivalent, which must be tested for the presence of a QTL in an AIC, is greater than in a simple cross. The genome-wide significance threshold must therefore rise. As an approximation, if QTL are located to 20-cM intervals in an F_2 but to 4-cM intervals in an AIC, then a genome-wide significance threshold of $-\log_{10}(p) = 3$ (say) would need to be increased to 3.7 in the advanced intercross. It is possible that a two-stage strategy could be derived in which QTL detection is first undertaken in a standard population with a lenient significance threshold with validation and fine mapping of these intervals occurring in an AIC as a second stage. The optimization of such a mapping strategy across two populations, which could also involve selective genotyping and phenotyping, merits study.

12 Near Isogenic Lines (NILs)

Near isogenic lines (NILs) are derived via repeated backcrossing of genetically distinct parental lines, most commonly with the aim of transferring a single chromosomal region of the donor parent into the genetic background of the recipient parent (reviewed by Kooke et al. [41]). Ultimately, over several generations of backcrossing, selection, and phenotyping, the interval containing the QTL should be eroded to something quite small. In practice, backcrossing beyond the second or third backcross is seldom carried out, the interval containing the QTL can remain large, and significant additional regions of donor parent genome can remain in the genetic background. The main use of NILs is that the effect of the QTL (and surrounding genome) can be characterized in much greater detail simply by phenotyping two lines rather than the whole, or a substantial part, of the mapping population. The power of NILs to detect a phenotypic effect of the QTL may not be substantially greater than from a similar size: if the QTL is the sole cause of genetic variation in the cross, power will be identical for identically-sized experiments. However, for a trait with a heritability of 50% of which the QTL accounts for 10%, the size of the mapping population would need to be 1.67 times greater than the experiment with NILs to get the same power.

Additionally, NIL pairs can be crossed, for example to fine-map a single QTL within the interval, or to combine two or more regions to investigate combined QTL effects. Collections of NILs that collectively capture all of the donor parent genome can be used for genetic mapping. Generation of such NIL populations, also termed chromosome segment substitution lines (CSSLs), can be aided via the use of molecular markers. Here, markers are used to both select for specific donor regions (foreground selection) and select against unwanted donor regions elsewhere in the genetic background (background selection), ultimately leading to an inbred population that can be used for genetic mapping. A recent example is the development of NIL populations in *Arabidopsis* (75 NILs [42]).

The backcrossing and purification required to develop NILs take time, leading to lag in their creation following QTL discovery by some other means. In mapping populations of inbred lines produced by selfing, inbreeding is seldom complete; the probability that an inbred line is fully homozygous after six generations of selfing is 0.03 for a species with 21 chromosome pairs and a total map length of 17 Morgans [43]. At this stage, the average line will contain 3.5 heterozygous tracts of chromosome and the total length of the genome that is heterozygous will be 27 cM on average. As a result, it is likely that individuals who are heterozygous, or families that are still segregating, for any tract can be found. Such heterogeneous inbred families (HIFs) can be used to rapidly create NILs through selfing [44]. Essentially the same approach was been used by Yamanaka et al. [45], for example, to fine map the *FTI* locus for soybean flowering time to a distance of 0.1 cM from the closest marker. In this case, a fully homozygous F₈ inbred line, aside from a 17 cM heterozygous tract around the *FTI* QTL was identified and 18 F₉ individuals

from the same line were selfed to create a population of $>1,006 F_{10}$ individuals that were used for mapping.

The utility of HIFs for power and precision is, of necessity, restricted to mapping populations produced by selfing rather than through DH. The probability of detecting HIFs for any particular QTL will depend on the generation of selfing. In the case of Yamanaka et al. [45], assuming the F_8 family originated from a single F_7 individual, there is a 0.0156 probability that the F_8 is segregating. Yamanaka tested 210 plants so the probability of *not* finding a segregating family is $(1-0.0156)^{210}$ or 3.7%: even with deep inbreeding, provided population sizes are modest, it is likely that HIFs will be found in bi-parental mapping populations and can be used to create NILs. This opens up the possibility of creating a tiled array of HIF's covering the genome.

13 Multi-Founder Populations

Within the last few years, and in parallel with the advent of association mapping for crops, more complex mapping populations have been advocated and used. These benefit from capturing increased levels of genetic diversity from the use of multiple founders. The use of multiple founders can also allow the incorporation of LD relationships among the founder into the analysis, with a potential gain in precision. These properties make multi-founder populations well suited to investigation of multiple traits, and are therefore used as community resources for genetic research. It is a feature of multi-founder populations that they are large: as a second-generation approach to mapping, it has been recognized that small populations are underpowered: a problem that is compounded in fine mapping where experimental methods to increase precision generally reduce power of QTL detection.

Multi-founder populations differ conceptually from bi-parental populations. Greater effort is required in their creation and they are generally designed to map multiple QTL for multiple traits. They are not a "use once and throw away" resource in the same way that bi-parental mapping populations can be. Choice of founders is important. The expectation is that inferences and discoveries in a multi-founder population will be applied to a wider population. An explicit understanding of the extent or range of this population is likely to give rise to the best choice of founders. For instance, selecting a set of lines to maximize global diversity is not the best strategy if the prime interest is to map traits for productivity in one particular agro-ecological environment. As with bi-parental mapping populations, there would be a risk, though reduced, of mapping QTL for loci for which the favorable allele is already fixed in the target environment. However, if the prime interest is in positional cloning, the best strategy may well be to select a set of founders that maximize species or crop diversity, though conditioned by the understanding that the lines that are produced will require phenotyping. This could cause problems in crosses between wild and cultivated forms, for example, or at least limit the range of traits that can be scored.

13.1 *Nested-Association Mapping*

Nested-association-mapping (NAM) populations, first proposed by Yu et al. [46] for the outcrossing species maize (*Zea mays*), are based on crossing multiple inbred lines to a single reference line then deriving multiple bi-parental sub-populations, either as DH lines or RILs. Originally designed for outcrossing species where LD decay is rapid (in maize, LD decays to background levels within ~1 kb), NAM combines the benefits of linkage mapping and association mapping: the linkage analysis within populations gives power to detect loci without the need for very high marker coverage, while the exploitation of more rapidly decaying LD across the multiple founders improves precision [46]. Protection against false positives arising from population structure is provided, in effect, by testing for association in the presence of linkage, an approach analogous to that of the QTDT [47, 48] in human genetics. NAM can capture high genetic variation while avoiding the complications of population structure, as usually found in GWAS panels.

The first NAM population was published in 2009, consisting of 25 maize inbred lines crossed to a single recurrent parent, resulting in 200 RILs per sub-population, and 5,000 RILs in total [49]. Since then, the maize NAM population has been extensively used for genetic analysis of multiple traits, including morphological, disease resistance, and metabolite phenotypes (e.g., [50–58]). Subsequently, NAM populations have been created in sorghum [59], as well as the inbreeding species, barley (*Hordeum vulgare*) [60] and wheat [61, 62]. While the classic NAM population design as exemplified by McMullen et al. [49] has invariably been used to date, related designs have been advocated. Simulation using empirical data in maize and *Arabidopsis* has shown that given a fixed total population size, power and reduction of false positives is optimized via employing designs such as diallel or factorial designs for outbred species [63].

For inbreeding species, where consideration was given to the effort needed to achieve the crosses required, double round robin design was found to be a good alternative [63]. Recently, an advanced backcross NAM (AB-NAM) population has been developed in barley that introgresses wild barley landraces into the exotic background [64]. The populations were developed by backcrossing 25 wild barley accessions to an elite barley cultivar. The lower proportion of wild genome in the recombinant lines makes phenotyping and mapping of loci in unadapted material easier.

NAM populations involve fewer crosses than alternative designs such as MAGIC (discussed below), and additional crosses can be added over time. Moreover, NAMs can emerge as a bi-product of other breeding and research activities. For example, the Wheat Improvement Strategic Programme (WISP) [65] is creating novel allohexaploid germplasm in wheat (synthetic wheat) by crossing tetraploid wheat with wild diploid goat grass (*Aegilops tauchii*). New synthetics are backcrossed to two elite lines and recombinant inbreds produced. This work was initiated purely for pre-breeding, but in essence has also created a NAM similar to that of Nice et al. [64]. A similar exercise in the WISP is producing lines from

backcrosses of landraces to elite wheat; also producing a NAM. Care must be taken with these emergent resources, however, to be aware of their statistical power. As they are not necessarily designed for mapping in the same way as the maize NAM, the numbers of lines created may be quite low. Power in mapping [66], as in life [67], is always worth thinking about.

Although multiple parental lines are involved in NAM, the creation of haplotype diversity is limited. With 26 lines at most 50 recombinant haplotypes can be created between two loci (out of 67,108,864 possible, assuming the parental loci are all difference). A greater limitation may be that these 50 will always involve the common parent: no novel haplotypes between the 25 unique parents are generated. Guo et al. [48] proposed that NAMs were created with two recurrent parents to avoid the emphasis on detection of QTL for which the recurrent and non-recurrent parents differ. This would create more haplotype diversity (100 recombinant haplotypes), but the potential for generating novel haplotype diversity is not as great as with MAGIC populations, as discussed below.

Where founders have been sequenced, lower-cost genotyping approaches (such as SNP arrays, GbS, or low-pass sequencing) in the progeny allows founder genotype to be projected onto the progeny. The prospect of sequencing the genomes of all individuals to medium-to-high coverage within a mapping population is now being realized. To date this has occurred in diversity and association-mapping panels, as this approach provides a wide survey of genetic diversity within a species. Examples include 3,000 rice accessions to $14\times$ sequencing depth [68] and 80 accessions to $\sim 15\times$ depth within the *Arabidopsis* 1,001 Genomes Project [69]. The continued development of sequencing technologies means that the application of genome re-sequencing to other types of plant mapping populations is inevitable in the near future.

13.2 *Heterogeneous Stock*

Mott et al. [70] proposed the use of an outbred mouse population (heterogeneous stock), created from eight inbred laboratory strains intercrossed for 60 generations, for fine mapping. Subsequently, Valdar et al. [71] mapped 843 QTL for multiple complex traits including aspects of behavior, which were located to 95% confidence intervals of ~ 2.8 Mb on average. Historically, development of heterogeneous stock in the mouse was instrumental in initiating the development of the first MAGIC populations in *Arabidopsis* and wheat. However, the mouse heterogeneous stock was not developed initially for mapping. Similar populations exist in crops, also not developed for trait mapping. In outbreeding species, including many crops, populations are often maintained in isolation for recurrent selection programs. For highly heritable traits, these can be used directly for mapping, though for traits with low heritability, clones (or inbred lines) would need to be extracted for phenotyping first. Similar populations exist in inbreeding species too. In wheat, for example, a French population was established with 60 founders, segregating for

genetic male sterility [72]. Provided seed is only ever harvested from the male sterile individuals, the population is maintained in a crossbred state. Population structure effects should be minimal, though there will be close familial relationships, so the risk of spurious marker trait associations should not be high. 1,000 inbred lines have been extracted from the French population after 12 generations of outcrossing. The population had been maintained as a bulk plot grown outside, so natural selection occurred, following the principles of Dynamic Management [72]. Thépot et al. [73] reported the first results from the French population, detecting 26 genomic regions under selection, of which six were associated with flowering time.

13.3 *MAGIC*

Following the success of the heterogeneous stock in mouse, we advocated the development of similar resources in crops – which we termed MAGIC populations [25]. These are in essence identical in construction to the mouse Collaborative Cross [74], with the exception that inbred lines are typically produced by selfing rather than by sib-mating (all that is possible in animals). Since then, MAGIC populations (reviewed by [75]) have been developed for many plant species (Table 1), including many of the world’s most important crops, e.g., rice [76, 77], wheat [78, 79], maize [49], and tomato [80]. The key characteristics of MAGIC populations are the use of multiple founders (typically eight) and multiple rounds of intercrossing, before the development of progeny for genetic mapping. With eight founders, there are 28 possible F_1 (2-way) crosses and 210 possible four-way crosses among unrelated F_1 s. There are then 315 possible ways of creating the eight-way crosses [79] (Fig. 1). Depending on the species, at the four-way stage and eight-way stage, the amount of crossing can become impractical and reduced numbers may be considered. In addition, progeny of four-way crosses are segregating, so there can be an advantage to replicating eight-way crosses with additional four-way parents. The design options and consequences for MAGIC populations have not been fully exhausted, though it is important to maintain a balance of contribution in lines of descent from each founder, to avoid introducing population structure into the population and produce as uniform a decay in LD across the genome as possible.

MAGIC populations afford a number of important benefits over the more commonly used bi-parental and/or association-mapping populations: (1) using multiple parent samples more genetic variation than in any traditional bi-parent cross. (2) The allele frequencies are balanced, because founders contribute equally to the population. (3) Dense, evenly distributed recombination sites provide considerable resolution for genetic analysis, genetic map construction, and gene isolation. MAGIC will work well in species where LD is extensive (such as inbreeding species like rice and wheat), and where LD mapping approaches may not give adequate precision, thus requiring more highly recombined resources. Combined

Table 1 Examples of published plant multi-founder populations

Population type	Species	Founder number	Population details	Genotype data	Reference
NAM	Maize	26	25 ILs crossed to 1 reference line, 5,000 RILs	1.5k SNPs	McMullen et al. [49]
NAM	Barley	26	25 wild barleys crossed to 1 reference line, 1,420 RILs	5.7k SNPs	Maurer et al. [60]
		26	25 wild barleys crossed to 1 reference line, 796 RILs	384 SNPs ^a	Nice et al. [64]
NAM	Bread wheat	11	10 varieties crossed to 1 reference line, 852 RILs	13.4k GbyS-derived SNPs	Bajgain et al. [61]
MAGIC	Bread wheat	4	2 rounds of intercrossing, 1,579 RILs	1,670 DArT/SSR/SNPs	Huang et al. [78]
		8	3 rounds of intercrossing, 1,000 RILs	82k SNP array	Mackay et al. [79]
MAGIC	Tomato	8	3 rounds of intercrossing, 397 RILs	1.5k SNP array	Pascual et al. [80]
MAGIC	Rice	8	ssp. <i>indica</i> . 3 rounds of intercrossing, 1,328 S7 RILs	GbS	Bandillo et al. [76]
		8	ssp. <i>indica</i> . 5 rounds of intercrossing		Bandillo et al. [76]
		8	ssp. <i>japonica</i> . 3 rounds of intercrossing, 500 S5 RILs		Bandillo et al. [76]
		16	4 rounds of intercrossing, ssp. <i>japonica</i> and <i>indica</i>		Bandillo et al. [76]
		4	ssp. <i>indica</i> . 2 rounds intercrossing, 271 RILs	6k SNP array	Meng et al. [77]
		4	ssp. <i>indica</i> . 2 rounds intercrossing, 268 RILs	6k SNP array	Meng et al. [77]
		8	ssp. <i>indica</i> . 3 rounds intercrossing, 531 RILs	6k SNP array	Meng et al. [77]
MAGIC	Barley	8	3 rounds intercrossing, 533 DH lines	4.5k SNPs	Sannemann et al. [90]
MAGIC	<i>Arabidopsis</i>	19	527 RILs	1.3k SNPs	Kover et al. [91]

(continued)

Table 1 (continued)

Population type	Species	Founder number	Population details	Genotype data	Reference
MAGIC	Maize	8	3 rounds of intercrossing, 1,636 RILs	54k SNPs	Dell'Acqua et al. [92]
AMPRIL	<i>Arabidopsis</i>	8	2 rounds of intercrossing, 532 RILs	321 SSR/SNPs	Huang et al. [87]

SNP single nucleotide polymorphism, *DArT* Diversity Array Technology, *SSR* simple sequence repeat, *RIL* recombinant inbred line, *IL* inbred line, *GbyS* genotyping-by-sequencing, *Arabidopsis thaliana*, *Barley Hordeum vulgare*, *Bread wheat Triticum aestivum*, *Maize Zea mays*, *Rice Oryza sativa*, *Tomato Solanum lycopersicum*, Genotype details refer to those listed in the original publication; some populations have been overlaid with additional genotype data subsequently

^aAdditionally, 4,022 SNPs and 263,531 sequence variants imputed into the population via additional genotyping of the parents

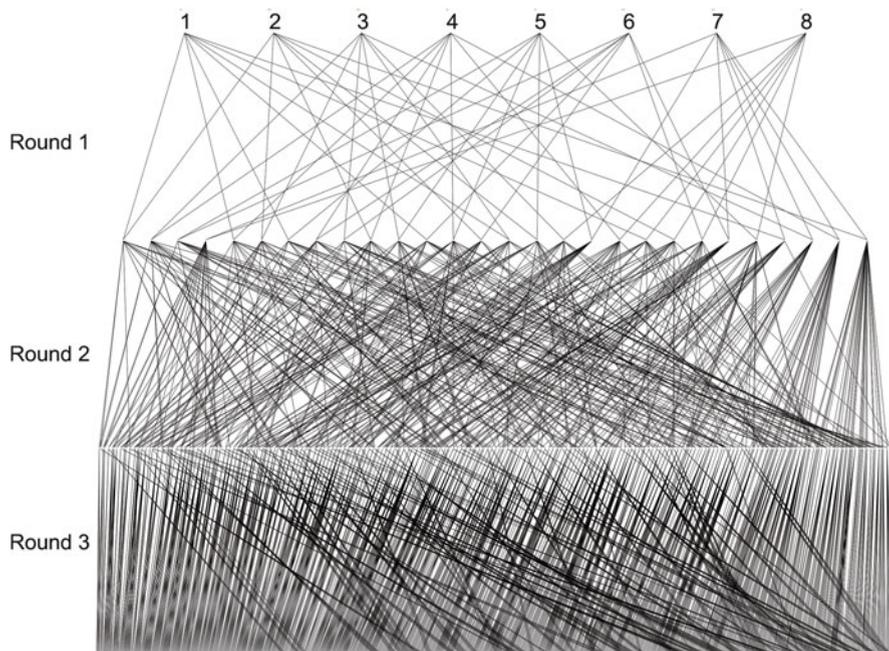


Fig. 1 A balanced eight-way MAGIC crossing scheme. Top line: eight founder parents. Second line: 28 two-way crosses resulting from all pairwise crosses (half-diallel) between the founders. Third line: 210 four-way crosses, resulting from intercrossing all unrelated two-ways. Fourth line: 315 eight-way crosses resulting from intercrossing all unrelated four-ways

with their suitability for the generation of high-density genetic maps, MAGIC populations are ideal community-based resources for crop improvement, fine mapping QTL, QTL \times environment and epistatic effects, and the anchoring of physical-genetic maps. However, a significant disadvantage of MAGIC populations, compared to say a bi-parental population, is the time they take to create. For example, the eight-way winter wheat MAGIC population created by Mackay et al. [79] took 3.5 years to complete: 1.5 years for the intercrossing, plus 2 additional years to reach the F₅ stage of inbreeding. Creation of DH lines from the final outcrossed population would reduce time, but a small number of DH lines are required from a large number of outcrossed individuals, unlike the case in a bi-parental population, where a large number of lines are required from a single individual (the F₁). This makes DH production to generate MAGIC populations prohibitively expensive for species like wheat.

An extension of MAGIC that has yet to be widely tested is the extraction of lines after further generations of outcrossing of the initial population. This was proposed by Mackay and Powell [25]. These lines would be more highly recombined than the initial set and therefore give finer location of QTL. High-density genotyping or sequencing of QTL containing regions identified in the initial population followed by selective phenotyping of recombinants should increase precision for only a modest cost. There is a limit to the number of additional generations of additional crossing that are worth carrying out: the reduction of linkage disequilibrium through crossing is countered by its increase as a result of finite population size together with loss of variation through drift. This process is under development in the 'NIAB Elite MAGIC' (eight-founders) and 'NIAB Diverse MAGIC' (16 founders) populations: an additional four cycles of crossing have been carried out and inbred lines are being created from the more advanced generation.

14 Analysis Approaches for MAGIC

There is a basic split of methods of analysis into those that regress traits onto individual markers and those that regress traits on probabilities of inheritance from founders. Within each of these categories there are then several variants. We give an outline of the three most common below. A fuller account and description are given in Verbyla et al. [81].

Standard statistical methods of analysis to test for marker-trait associations, such as a t-test, do not take into account the method of construction of MAGIC populations, and will result in an increased frequency of false-positive results. A typical MAGIC population is constructed from a set of funnel crosses in which founders are crossed in different orders. For an eight-founder population, there are 315 possible ways in which all founders can be combined. For any polygenic trait, the additive genetic variance expressed between funnels is expected to be 1/8th of the whole. In addition, (1) replicate crosses may be made when selecting parents from four-way and subsequent segregating generations, (2) replicate individuals

within the final generation of crosses may be selected to initiate inbreeding, and (3) more than one inbred line may be derived from each outcrossed individual. These nested relationships among the resulting lines must be taken into account during analysis. For example, we have used gene dropping, as implemented in the bespoke software GeneDrop [82] (available from <http://www.niab.com/>), to simulate a quantitative trait segregating in the “NIAB Elite MAGIC” population. We simulated an eight-founder population with an average of six inbred lines per funnel; 10,000 pairs of unlinked SNPs were tested for association. In the absence of correction for family structure, false-positive rates were 0.1678, 0.0322, 0.0056, and 0.0009 at nominal significance thresholds of 0.1, 0.01, 0.001, and 0.0001, respectively: a considerable increase over expected, particularly at more stringent significance thresholds. Incorporating family structure, the corresponding false-positive rates were 0.1056, 0.0126, 0.0011, and 0.0000: much closer to the nominal significance levels.

The hierarchical structure described above is most easily taken into account in a mixed model incorporating random effects (i.e., variance components) for differences at each level. In a typical eight-founder population, this would include terms for differences between funnels, between replicate crosses within funnels, and between replicate plants within replicate crosses. The marker effects of greatest interest are included as fixed effects. For bi-allelic markers a simple additive model is most powerful, in a one degree of freedom (df) test in which the trait is regressed on the gene dosage (0, 1, 2 in a diploid), of an arbitrarily chosen reference allele. A two df test can also be used in which the three genotype classes are treated separately. However, the heterozygous class is usually rare and, even for dominant QTL, the two df test generally has reduced power. The effect of the heterozygotes can readily be examined after the initial scan, however. Multi-allelic loci are best tested for linkage in a (no. of alleles – 1) test.

In this model, there is no requirement to estimate and incorporate into the analysis a genetic relationship matrix among SSD lines in the manner required in association mapping. An advantage of this approach is therefore that it is easy to implement in almost all standard statistical packages. In R [83] for example, the lmer package [84] can be used to incorporate the desired random effects. As a result, modelling multiple markers, their interactions, and other covariates is also straightforward.

Incorporation of a marker-based relationship matrix is also possible, however, and has been used in mapping within a rice MAGIC population [76]. It has the advantage that relationships among the founder lines are also taken into account, though computational ease and simplicity is reduced.

Single-marker methods of analysis should always be included: they are quick, flexible, and robust to genotype error, which will generally reduce power but not increase the false-positive rate. Alternative methods of analysis use the marker data to estimate probabilities of identity by descent between each RIL and each founder at all selected locations over the genome. Ideally, these probabilities will be one or zero, indicating that a particular location in a line is known to have originated with certainty from one of the founders. In the worst case, these probabilities would be

1/(number of founders). Software to calculate these probabilities is available in r/qtl [5], mpMap [85], RABBIT [86], and HAPPY [70]. The first three packages require the pedigree of each line (with reference to the founders) to be known, whereas the latter does not. We are not aware of an independent comprehensive analysis of the absolute and relative accuracies of these methods, while also taking into account their availability, reliability, and ease of use.

Once calculated, traits can be regressed on each founder probability to give a test with 1 df for a QTL allele carried by that founder. Since there would be (no. of founders -1) such independent tests, a multiple regression is carried out on the identity by descent (IBD) values to give a single test with (no. of founders -1) degrees of freedom. To achieve this, one of the founder probabilities (it doesn't matter which) is dropped. Just as for single-marker analyses described above, the hierarchical population structure of the MAGIC populations should still be taken into account. HAPPY does not do this, but estimates empirical significance thresholds through a resampling procedure. IBD probabilities can also be calculated for locations between markers. These too, can be used for analysis. This is analogous to interval mapping in bi-parental crosses. IBD methods are generally restricted to additive models: with eight founders there are 28 heterozygous combinations, so the locus specific test for association would require 35 df, assuming all heterozygous classes were represented. The IBD approach to analyzing MAGIC populations allows each founder to carry separate QTL alleles. This will be an advantage over the single (bi-allelic) marker approach in some circumstances. This depends on the true number of QTL alleles, the distribution of their effects, the pattern of LD between marker and QTL alleles, the accuracy of IBD determination, the accuracy of genotyping (which will disproportionately affect IBD probability estimation), and marker density. It is possible that in the near future, IBD methods may be superseded as methods of genotyping and sequencing result in marker densities approaching the limit of capturing all variants segregating in the population (though this will increase the problem of multiple-testing). Our best advice would be to try an IBD-based method and a single-marker method. If results agree, all well and good (and in our experience, they usually do). Lack of agreement should be explored further to establish the cause. More complex models (reviewed by Verbyla et al. [81]) involve approaches analogous to composite interval mapping, Bayesian methods, and can fit multiple QTL models with simultaneous analysis of phenotypes. Sadly, most methods are currently not easily accessible to the non-statistician or data analyst.

14.1 *Arabidopsis* Multi-Parent Recombinant Inbred Line (AMPRIL)

The *Arabidopsis* multi-parent recombinant inbred line (AMPRIL) population described by Huang et al. [87] was developed from eight inbred *Arabidopsis*

accessions from diverse geographical origins. Four unrelated F_1 combinations were made among the eight founders. These F_1 s were crossed in a diallel to give six four-way crosses (pooling reciprocals), which were then selfed to produce 532 inbred lines. This pattern of construction is similar to that for MAGIC, but required fewer crosses and generations to create than the equivalent eight-founder MAGIC population. The comments above about MAGIC populations therefore apply to AMPRIL too. For an equivalent population size, a MAGIC population will provide more resolution than AMPRIL, although AMPRIL would be quicker to create.

14.2 Linked or Multiple Mapping Populations

In principle, any set of mapping populations can be analyzed simultaneously to detect QTL and, by increasing sample size, to increase power and precision. If links between populations can be made, power may be increased further by incorporating information on linkage disequilibrium across populations in addition to linkage information within populations. This has resulted in the use of lines derived from various sets of linked crosses, such as from diallels [88], though these links are not an absolute requirement. The focus of these approaches has been largely on detection of QTL in different genetic backgrounds and on epistasis rather than primarily on improving precision, though that should be a consequence of increasing population size. A recent example, with references to earlier work, is that of Han et al. [89].

15 Conclusion and Outlook

Precision mapping requires that marker-trait associations are tested in populations in which large numbers of recombinations have occurred. To achieve this goal, there are two broad approaches: increase population size and increase the number of generations of crossing. The methods described in this review attempt one or both of these. We favor approaches such as MAGIC and AMPRIL, since these focus explicitly on increasing the amount of recombination that occurs within the population. This bias may be because our own background is of working with inbred crops where LD generally decays quite slowly in collections of lines and the number of elite cultivars available is limited. Consequently, the power and precision of association mapping may be limited. In contrast, approaches that rely on linked sets of crosses, such as NAM, may be better suited to outcrossed species such as maize (where the limitation can be that LD decays too quickly in association-mapping panels, so power is limited but precision is increased). In these circumstances, experimental populations may be required to increase power as much as to increase precision. An equivalent way of viewing the choice would be that MAGIC populations are better for fine mapping in germplasm of immediate relevance to

breeders' elite germplasm, but NAM is better for progression towards gene discovery and positional cloning, since greater diversity can be captured in very diverse germplasm and the use of linkage in addition to linkage disequilibrium protects against loss of power for QTL detection.

Whatever approach is followed, the days of mapping QTL in small populations must come to an end. In our own work in MAGIC wheat populations, we started with a target of developing 1,000 lines per population: that number now looks on the low side.

Acknowledgements JC and IM were partially funded by grants from the Biotechnology and Biological Sciences Research Council (BB/M008908/1, BB/M011666/1 and BB/L011700/1) and the Agriculture and Horticulture Development Board (RD2200003).

Glossary

Advanced inter-cross (AIC) A bi-parental population, in which founders have been intercrossed for two or more generations prior to the production of inbred lines.

Doubled haploid (DH) A genotype formed when haploid cells undergo chromosome doubling.

Genomic selection (GS) A form of marker-assisted selection in which genetic markers are combined with phenotypic data to estimate breeding values in the absence of precise knowledge of where specific genes are located.

Genome wide association scan (GWAS) Method for genetic mapping using a collection of varieties or landraces with phenotypic and genome-wide genotypic datasets.

Linkage disequilibrium (LD) The non-random association of alleles at separate loci located on the same chromosome.

Multiparent advanced generation inter cross (MAGIC) population A multi-founder population created by intercrossing the founders over multiple generations in a balanced crossing scheme, prior to the production of inbred lines.

Nested association mapping (NAM) population A multi-founder population created by generating multiple bi-parental inbred populations, each of which contains a common founder.

Quantitative trait locus (QTL) A polymorphic site contributing to the genetic variability of a quantitative trait.

Recombinant inbred line (RIL) A population developed by single seed descent from the F₂ generation.

References

1. Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8:552
2. Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. *Theor Appl Genet* 67:35–43
3. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721–726
4. Paux E, Sourdille P, Mackay I, Feuillet C (2012) Sequence-based marker development in wheat: advances and applications to breeding. *Biotechnol Adv* 30:1071–1088
5. Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890
6. Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* 363:557–572
7. Beavis WD (1998) QTL analysis: power, precision, and accuracy. In: Paterson AH (ed) *Molecular dissection of complex traits*. CRC Press, Boca Raton, pp 145–173
8. Kearsley MJ, Farquhar AG (1998) QTL analysis in plants; where are we now? *Heredity* 80:137–142
9. Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci* 48:1649–1664
10. Monna L, Lin HX, Kojima S, Sasaki T, Yano M (2002) Genetic dissection of a genomic region for a quantitative trait locus, *Hd3*, into two loci, *Hd3a* and *Hd3b*, controlling heading date in rice. *Theor Appl Genet* 104:722–778
11. Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. *Crop Sci* 55(1):12
12. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
13. Mackay I, Ober E, Hickey J (2015) GplusE: beyond genomic selection. *Food Energy Secur* 4:25–35
14. Bortesi L, Fischer R (2015) The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol Adv* 33:41–52
15. Hickey JM, Bruce C, Whitelaw A, Gorjanc G (2016) Promotion of alleles by genome editing in livestock breeding programmes. *J Anim Breed Genet* 133:83–84
16. Law CN, Worland AJ, Giorgi B (1976) The genetic control of ear-emergence time by chromosome 5A and 5D of wheat. *Heredity* 36:49–58
17. Price AH (2006) Believe it or not, QTLs are accurate! *Trends Plant Sci* 11:213–216
18. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376
19. Nature Genetics Editorial Board (2005) Framework for a fully powered risk engine. *Nat Genet* 37:1153
20. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9:356–369
21. Hall D, Tegström C, Ingvarsson PK (2010) Using association mapping to dissect the genetic basis of complex traits in plants. *Brief Funct Genomics* 9:157–165
22. Bentley AR, Scutari M, Gosman N, Faure S, Bedford F, Howell P, Cockram J, Rose GA, Barber T, Irigoyen J, Horsnell R, Pumfrey C, Winnie E, Schacht J, Beauchêne K, Praud S, Greenland A, Balding D, Mackay IJ (2014) Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. *Theor Appl Genet* 127:2619–2633

23. Cockram J, White J, Zuluaga DL, Smith D, Comadran J et al (2010) Genome-wide association mapping to candidate polymorphism resolution in the un-sequenced barley genome. *Proc Natl Acad Sci U S A* 107:21611–21616
24. Waugh R, Marshall D, Thomas B, Comadran J, Russell J, Close T, Stein N, Hayes P, Muehlbauer G, Cockram J, O’Sullivan D, Mackay I, Flavell A, Agoueb A, Barley CAP, Ramsay L (2010) Whole-genome association mapping in elite inbred crop varieties. *Genome* 53:967–972
25. Mackay IJ, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
26. MacArthur D (2012) Methods: face up to false positives. *Nature* 487:427–428
27. Highfill CA, Reeves GA, Macdonald SJ (2016) Genetic analysis of variation in lifespan using a multiparental advanced intercross Drosophila mapping population. *BMC Genet* 17(1):113
28. Cockram J, White J, Leigh FJ, Lea VJ et al (2008) Association mapping of partitioning loci in barley (*Hordeum vulgare* ssp. *vulgare* L.) *BMC Genet* 9:16
29. Darvasi A, Soller M (1995) Advanced intercross lines, and experimental population for fine genetic mapping. *Genetics* 141:1199–1207
30. Ma J, Wingen LU, Orford S, Fenwick P, Wang J, Griffiths S (2015) Using the UK reference population Avalon x Cadenza as a platform to compare breeding strategies in elite Western European bread wheat. *Mol Breed* 35:70
31. Bentley AR, Jensen EF, Mackay IJ, Hönicka H, Fladung M, Hori K, Yano M, Mullet JE, Armstead IP, Hayes C, Thorogood D, Lovatt A, Morris R, Pullen N, Mutasa-Göttgens E, Cockram J (2013) Genomics and breeding for climate-resilient crops (ed Kole C) volume II target traits chapter 1. Flowering time. Springer, Berlin
32. Bentley A, Mackay I (2016) Advances in wheat breeding techniques. In: Langridge P (ed) Achieving sustainable cultivation of wheat. Burleigh Dodds Science Publishing Ltd., Cambridge
33. Zou C, Wang P, Xu Y (2016) Bulk sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol J* 14:1941–1955
34. Routaboul J-M, Dubos C, Beck G, Marquis C, Bidzinski P, Loudet O, Lepiniec L (2012) Metabolite profiling and quantitative genetics of natural variation for flavonoids in *Arabidopsis*. *J Exp Bot* 63:3749–3764
35. Ries D, Holtgräwe D, Viehöver P, Weisshaar B (2016) Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* 17:236
36. Hill WG (1998) A note on the theory of artificial selection in finite populations and application to QTL detection by bulk segregant analysis. *Genet Res* 72:55–58
37. Mackay IJ, Caligari PDS (2000) Efficiencies in F₂ and backcross generations for bulked segregant analysis using dominant markers. *Crop Sci* 40:626–630
38. Fitz Gerald JN, Carlson AL, Smith E, Maloof JN, Weigel D, Chory J, Borevitz JO, Swanson RJ (2014) New *Arabidopsis* advanced intercross recombinant inbred lines reveal female control of nonrandom mating. *Plant Physiol* 165:175–185
39. Balint-Kurti PJ, Wisser R, Zwonitzer JC (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for gray leaf spot resistance in maize. *Crop Sci* 48:1696–1704
40. Balint-Kurti PJ, Zwonitzer J, Wisser R (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for grey leaf spot resistance in maize. *Crop Sci* 48:1696–1703
41. Kooke R, Wijker E, Keurentjes JJ (2012) Backcross populations and near isogenic lines. *Methods Mol Biol* 871:3–16
42. Fletcher RS, Mullen JL, Yoder S, Bauerle WL, Reuning G, Sen S, Meyer E, Juenger TE, McKay JK (2013) Development of a next-generation NIL library in *Arabidopsis thaliana* for dissecting complex traits. *BMC Genomics* 14:655
43. Gale JS (1980) Population genetics. Blackie and Son, Glasgow and London

44. Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet* 95:1005–1011
45. Yamanaka N, Watanabe S, Toda K, Hayashi M, Fuchigami H, Takahashi R, Harada K (2005) Fine mapping of the *FT1* locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. *Theor Appl Genet* 110:634–639
46. Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551
47. Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–292
48. Guo B, Sleper DA, Beavis WD (2010) Nested association mapping for identification of functional markers. *Genetics* 186:373–383
49. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Lu H et al (2009) Genetic properties of a maize nested association mapping population. *Science* 325:539–551
50. Brown PJ, Upadaya N, Mahone GS, Tian F, Bradbury PJ et al (2011) Distinct genetic architectures for male and female inflorescence traits of maize. *PLoS Genet* 7:e1002383
51. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
52. Hung H-Y, Shannon LM, Tian F, Bradbury PJ, Chen C et al (2012) *ZmCCT* and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. *Proc Natl Acad Sci U S A* 109:E1913–E1921
53. Kump KL, Bradbury PJ, Wissner RJ, Buckler ES, Belcher AR et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163–168
54. Peiffer JA, Flint-Garcia SA, De Leon N, McMullen MD, Kaeppler SM et al (2013) The genetic architecture of maize stalk strength. *PLoS One* 8:e67066
55. Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z et al (2014) The genetic architecture of maize height. *Genetics* 196:1337–1356
56. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc Natl Acad Sci U S A* 108:6893–6898
57. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
58. Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES (2014) Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* 10:e1004845
59. Jordan D, Mace E, Cruickshank A, Hunt C, Henzell R (2011) Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Sci* 51:1444–1457
60. Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, Schumann E, Killian B, Reif JC, Pillen K (2015) Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16:290
61. Bajgain P, Rouse MN, Tsilo TJ, Macharia GK, Bhavani S, Jin Y, Anderson JA (2016) Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. *PLoS One* 11:e0155760
62. Wingen LU, West C, Leverington-Waite M, Collier S, Orford S et al (2017) Wheat landrace genome diversity. *Genetics* 205:1657–1676
63. Stich B (2009) Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*. *Genetics* 183:1525–1534
64. Nice LM, Steffenson BJ, Brown-Guedira GL, Akhunov ED, Liu C, Kono TJY, Morrell PL, Blake TK, Horsley RD, Smith KP, Meuhlbauer GJ (2016) Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild x cultivated barley. *Genetics* 203:1453–1467
65. Moore G (2015) Strategic pre-breeding for wheat improvement. *Nat Plants* 1:15018

66. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202
67. Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychol Bull* 76:105
68. 3000 Rice Genomes Project (2014) The 3000 rice genomes project. *Gigascience* 3:7
69. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43:956–963
70. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci U S A* 97:12649–12654
71. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R, Flint J (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* 38:879–887
72. Goldringer I, Enjalbert J, David J, Paillard S, Pham JL et al (2001) Dynamic management of genetic resources: a 13-year experiment on wheat. In: Cooper HD, Spillane C, Hodgkin T (eds) *Broadening the genetic base of crop production*. CABI, Wallingford, pp 245–260
73. Thépot S, Restoux G, Goldringer I, Gouache D, Mackay I, Enjalbert J (2015) Efficiently tracking selection in a multiparental population: the case of earliness in wheat. *Genetics* 199:609–623
74. The Complex Trait Consortium (2002) The collaborative cross, a community resource for the genetic analysis of complex traits. *Nat Genet* 36:1133–1137
75. Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P, Leung H, Varshney RK, Cavanagh CR (2015) MAGIC populations in crops: current status and future prospects. *Theor Appl Genet* 128:999–1017
76. Bandillo N, Raghavan C, Muyca PA, Sevilla MAL, Lobina IT (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetic research and breeding. *Rice* 6:11
77. Meng L, Guo L, Ponce K, Zhao X, Ye G (2016) Characterization of three *indica* rice multiparent advanced generation intercross (MAGIC) populations for quantitative trait loci identification. *Plant Genome* 9(2).
78. Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis of wheat. *Plant Biotechnol J* 10:826–839
79. Mackay I, Bansept-Basler P, Barber T, Bentley AR, Cockram J et al (2014) An eight-parent multiparent advanced generation intercross population for winter-sown wheat: creation, properties and validation. *G3 (Bethesda)* 4:1603–1610
80. Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet JP, Le QH, Chauchard B, Verschave P, Causse M (2015) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol J* 13:565–577
81. Verbyla AP, George AW, Cavanagh CR, Verbyla KL (2014) Whole-genome QTL analysis for MAGIC. *Theor Appl Genet* 127:1753–1770
82. Ladejobi O, Elderfield J, Gardner KA, Gaynor RC, Hickey J, Hibberd JM, Mackay IJ, Bentley AR (2016) Maximizing the potential of multi-parental crop populations. *App Transl Genom* 11:9–17
83. R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: <https://www.R-project.org/>
84. Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48
85. Huang BE, George AW (2011) R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. *Bioinformatics* 27:727–729

86. Zheng C, Boer MP, van Eeuwijk F (2015) Reconstruction of genome ancestry blocks in multiparental populations. *Genetics* 200:1073–1087
87. Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proc Natl Acad Sci U S A* 108:4488–4493
88. Rebai A, Goffinet B (1993) Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theor Appl Genet* 86:1014–1022
89. Han S, Utz HF, Liu W, Schrag TA, Stange M, Würschum T, Miedaner T, Bauer E, Schön CC, Melchinger AE (2016) Choice of models for QTL mapping with multiple families and design of the training set for prediction of Fusarium resistance traits in maize. *Theor Appl Genet* 129:431–444
90. Sannemann W, Huang BE, Mathew B, Léon J (2015) Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. *Mol Breed* 35:86
91. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet* 5:e7
92. Dell'Acqua M, Gatti DM, Pea G, Cattonaro F, Coppens F, Magris G, Hlaing AL, Aung HH, Nelissen H, Baute J, Frascaroli E, Churchill GA, Inzé D, Morgante M, Pé ME (2015) Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol* 16:167