Theor Appl Genet (2015) 128:999–1017
DOI 10.1007/s00122-015-2506-0

REVIEW

MAGIC populations in crops: current status and future prospects

B. Emma Huang[1] · Klara L. Verbyla[2] · Arunas P. Verbyla[3] · Chitra Raghavan[4] · Vikas K. Singh[5] · Pooran Gaur[5] · Hei Leung[4] · Rajeev K. Varshney[5,6] · Colin R. Cavanagh[7]

Received: 17 October 2014 / Accepted: 20 March 2015 / Published online: 9 April 2015

Abstract
Key message **MAGIC populations present novel challenges and opportunities in crops due to their complex pedigree structure. They offer great potential both for dissecting genomic structure and for improving breeding populations.**
Abstract The past decade has seen the rise of multiparental populations as a study design offering great advantages for genetic studies in plants. The genetic diversity of multiple parents, recombined over several generations, generates a genetic resource population with large phenotypic diversity suitable for high-resolution trait mapping. While there are many variations on the general design, this review focuses on populations where the parents have all been inter-mated, typically termed Multi-parent Advanced Generation Intercrosses (MAGIC). Such populations have already been created in model animals and plants, and are emerging in many crop species. However, there has been little consideration of the full range of factors which create novel challenges for design and analysis in these populations. We will present brief descriptions of large MAGIC crop studies currently in progress to motivate discussion of population construction, efficient experimental design, and genetic analysis in these populations. In addition, we will highlight some recent achievements and discuss the opportunities and advantages to exploit the unique structure of these resources post-QTL analysis for gene discovery.

Communicated by H. H. Geiger.

✉ B. Emma Huang
emma.huang@csiro.au

[1] Digital Productivity and Agriculture Flagships, CSIRO, Dutton Park, QLD 4102, Australia

[2] Digital Productivity and Agriculture Flagships, CSIRO, Canberra, ACT 2601, Australia

[3] Digital Productivity and Agriculture Flagships, CSIRO, Atherton, QLD 4883, Australia

[4] Plant Breeding, Genetics and Biotechnology Division, International Rice Research Institute, 1301 Manila, Philippines

[5] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502324, India

[6] School of Plant Biology and Institute of Agriculture, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

[7] Agriculture Flagship, CSIRO, Canberra, ACT 2601, Australia

Introduction

A major advantage for researchers in plant and animal genetics lies in the ability to create experimental populations. Such populations mix well-characterized founder genomes in controlled pedigrees, and facilitate the investigation of both the genome itself and its relationship with traits and the environment. Traditional experimental populations combine the genomes of two parents with contrasting phenotypes to identify regions of the genome affecting the trait. However, each of these populations captures only a small snapshot of the factors affecting the trait due to the narrow genetic base—it is only possible to detect those genomic regions which differ between the two founders, and all alleles occur with high frequency in the population. The alternative of association mapping takes a panoramic view of the whole population by sampling distantly related individuals. It hence captures

far greater diversity, but requires very large samples to have sufficient power to detect genomic regions of interest, and hence may have difficulty detecting rare alleles of importance.

The weaknesses of existing designs have led to a new type of complex experimental design intermediate to biparental and association mapping designs in terms of power, diversity, and resolution. Multi-parent Advanced Generation InterCrosses (MAGIC) inter-mate multiple inbred founders for several generations prior to creating inbred lines, resulting in a diverse population whose genomes are fine-scale mosaics of contributions from all founders. Similar to biparental populations, alleles occur at relatively high frequencies due to the limited number of founders, but the population encapsulates much higher diversity in polymorphisms. While a MAGIC population requires greater initial investment in capability and time than a biparental, careful selection of founders allows its generalizability to the wider breeding population and ensures relevance as a long-term genetic resource panel.

The first multiparental inter-mated population was the Collaborative Cross (CC, Complex Trait Consortium 2004) in mice, but the design has since had wide uptake across a variety of species. The CC combined the genomes of eight inbred strains together through multiple intercrosses, and then created inbred lines through sibling mating. Concurrently, the Diversity Outbred (DO) population was developed as a related heterogeneous stock (HS) population with the same eight progenitors (Collaborative Cross Consortium 2012). These two populations form an extensive genetic resource for mouse which has been utilized for mapping and identification of candidate genes for serum cholesterol and coat color traits (Svenson et al. 2012; Ram et al. 2014). A similar design was used to create the Drosophila Synthetic Population Resource (King et al. 2012a). In plants, there has been more variety in the features of different populations, due to many differences between organisms including genome size and complexity, history of the species, and resources and technologies available. MAGIC populations have already been created in *Arabidopsis thaliana* (Kover et al. 2009), wheat (Huang et al. 2012; Mackay et al. 2014), rice (Bandillo et al. 2013) and are underway in chickpea (Gaur et al. 2012) and a variety of other crops.

While much of the motivation and challenges inherent in these populations are common to different organisms, moving from model organisms to crops generally poses a new set of challenges and questions of interest. Reference sequence data may not be available, polyploidy is common, the physical scale of the experiments will often be much larger, and phenotypic data are often collected outside controlled conditions. In this review, we describe some of the largest MAGIC studies currently in progress in crops, using them to motivate discussion of design and analysis issues in MAGIC. We conclude with lessons learned from existing populations, highlighting some recent achievements and future directions for multiparental genetics.

Efficient experimental design

Careful consideration of design prior to initiating population development helps to ensure not only the novelty of a population, but also its ability to answer practical questions of interest. For MAGIC populations this is of particular importance given the complexity of the design, the time investment required for development, and the number of factors which eventually impact the power, diversity, and resolution of the progeny. Hence, the objectives for the population need to be clearly defined before embarking on population development. We will touch both on the factors which pertain to the pedigree and how the founder lines are inter-mated, as well as additional considerations which may improve the efficiency of the study design. Figure 1 depicts the stages of population development which are described in further detail below.

Founder selection Prior to initiating population development, founder lines must be chosen (Fig. 1a). This may be based on genetic and/or phenotypic diversity, either in a constrained set of material (e.g., elite cultivars, geographical adaptation) or material of more diverse origins (worldwide germplasm collections, distant relatives). Achieving an optimal level of genetic diversity is not a simple task. Use of landraces as founders may introduce greater diversity, but simultaneously reduce the generalizability to the current breeding populations. In addition, genetic incompatibility in some species can cause a large reduction in the number of progeny that may be derived from specific crosses. Variety-specific gross chromosomal differences such as rearrangements or alien/wild introgressions may also affect the production of the final population and its use for genetic mapping. On the other hand, narrow genetic diversity can be problematic for estimating founder probabilities (see Sect. "Genetic analysis") and prevent researchers from fully exploiting the potential of their populations.

In addition to genetic diversity, the phenotypic diversity must be carefully managed to produce a resource which is also practical. Consideration of traits such as flowering time in the founders will avoid segregation for undesirable values in the progeny which will affect not only subsequent phenotypic evaluation, but also have practical impact on making the crosses. Ultimately, the selection of founders will be one of the most important design considerations and

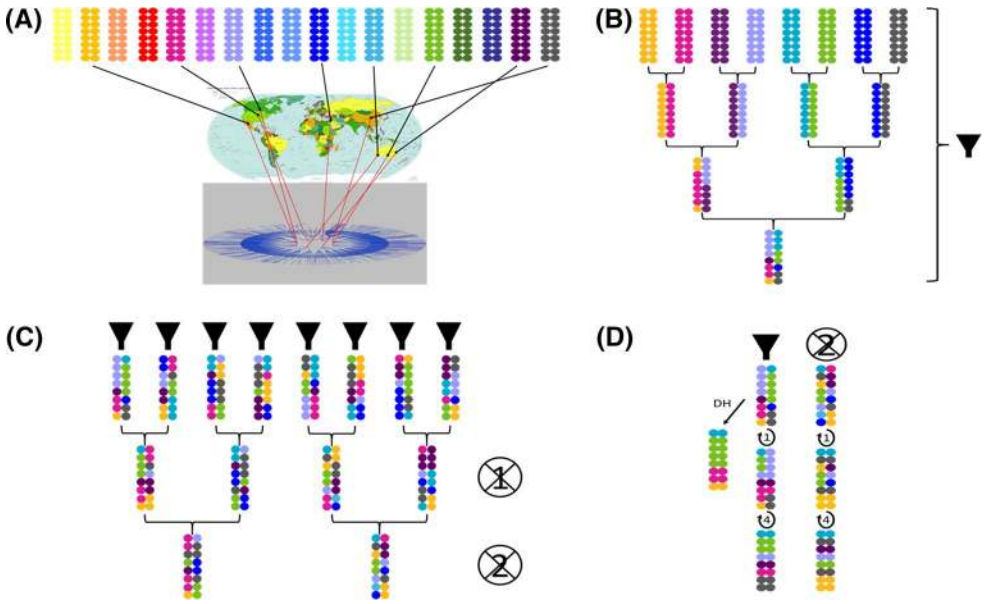

Fig. 1 Stages of MAGIC population development design for eight founders **a** selection of founders based on geographic, genetic, phenotypic diversity. The maternal pedigree tree is presented at the *bottom* for an eight-way MAGIC population with each *ring* representing a subsequent generation; **b** mixing of parents together in predefined patterns, or funnels (denoted by *symbol* on *right*); **c** intercrossing of individuals (generations denoted by *number* within *crossed circle*) derived from different funnels for additional generations; **d** selfing (generations denoted by number within *circular arrow*) or double haploidization of individuals either directly from funnels or after advanced intercrossing to form *inbred lines* (color figure online)

will depend heavily on the goals for the population. More diverse founder sets may provide biological insight into a wide variety of traits; however, founders selected based on relevance to a breeding program for specific traits may result in a MAGIC population which more quickly translates into superior breeding lines. In particular, in breeding it may be desirable to focus on capturing the allelic diversity present within the breeding program across market segments, and ensuring the population size is adequate to ensure sufficient progeny expected to fit each market segment.

Mixing In the first stage of population development, multiple parents are intercrossed to form a broad genetic base (Fig. 1b). This was inspired by the heterogeneous stock (HS), proposed by McClearn et al. (1970) and taken up by Demarest et al. (2001), which goes on to create an outbred population derived from multiple parents. In this stage, the inbred founders are paired off and inter-mated in a prescribed order for each line, known as a *funnel*. If each recombinant inbred line (RIL) is the product of a 2^n-way cross, then the mixing stage will require n generations. The result of this stage is a set of lines whose genomes comprised contributions from each of the founders. Broman (2005) showed that the composition of these contributions depends on the funnel structure; hence, the design and selection of funnels used in the cross will impact on the eventual genetic makeup of the population.

Greater variety in the types of funnels generated will ensure greater robustness to accounting for factors such as maternal effects, population structure, and segregation distortion; however, it will require greater investment in terms of cost and time. Simulation may provide more insight into balancing these factors (see Sect. "In silico experiments"). A further concern in generating a limited number of funnels is the level of relatedness due to shared recombination events. If at any generation the number of individuals is small, individuals derived from them will be related to each other, reducing the genetic diversity. This relatedness can also create population structure which will bias analyses such as linkage map construction and QTL mapping if not appropriately accounted for.

Advanced intercrossing In the second stage (Fig. 1c), the mixed lines from different funnels are randomly and sequentially intercrossed as in the advanced intercross (AIC) proposed by Darvasi and Soller (1995). The main goal of this intercrossing is to increase the number of recombinations in the population. Yamamoto et al. (2014) performed a simulation study to consider the effect of different numbers of generations of intercrossing on genome structure. They concluded that at least six cycles of intercrossing were required for large improvements in QTL mapping power. Lines undergoing differing generations of advanced intercrossing may be combined together in the analysis; however, evidence of population structure

should be investigated (e.g., using STRUCTURE, Pritchard et al. 2000) to ensure that the intercrossing has not introduced differences in structure between the subsets of the population.

Inbreeding In the third stage, the individuals resulting from the advanced intercrossing stage are progressed to create homozygous individuals (Fig. 1d). RILs in plants can be created via single seed descent (Goulden 1939; Brim 1966; Bailey 1971) or doubled haploid production (Blakeslee et al. 1922; Maluszynski et al. 2003; see Forster et al. 2007 for a review of methods). While doubled haploid production is often faster, the multiple generations of selfing will introduce additional recombination, albeit less than during the mixing and advanced intercrossing stages.

The progeny of the population will not be fully inbred in practice except for double haploid lines. This residual heterozygosity can be both useful and problematic. In genotyping, it may cause issues due to the inability to distinguish heterozygotes for some markers, particularly for polyploids (Cavanagh et al. 2013) and genotyping-by-sequencing (GBS) approaches (Elshire et al. 2011). In data analysis, it may cause issues by violating the simplifying assumption of full inbreeding (Broman 2005), and should ideally be addressed as in Broman (2012), which derives genotype probabilities for individuals at intermediate generations with substantial heterozygosity. In many of the populations discussed here, plants have been self-pollinated for five or more generations, so the expected level of heterozygosity in the genome is less than 3 %. In general to develop MAGIC populations a minimum of 8 crop seasons is required to reach at least the S5 generation.

In silico experiments

Simulation is an important tool for understanding the potential of different designs, for comparing methodology, and for developing guidelines for future studies. A number of simulation packages are now publically available for multiparental designs, with varying levels of flexibility. R/qtl (Broman et al. 2003) and R/mpMap (Huang and George 2011) allow the specification of genotypes for founders, and descend them through a pedigree. While R/mpMap allows greater flexibility in pedigree definition, R/qtl is more flexible in modeling genetic processes such as crossover interference. A more recent addition, AlphaMPSim (Hickey et al. 2014), generates founder genotypes through a coalescent model prior to gene dropping, and is built to efficiently generate data up to full sequence.

With regard to simulating different-sized studies, it is important to consider both the total population size and the size of the subset which may be phenotyped for an individual trait. In many crops, replicates of genotypes are necessary in field trials to estimate environmental variability, and hence even if a population of 1000 lines exists it may not be feasible to phenotype them all in a single study. Thus, consideration of the total population size may be more affected by the practicalities of its maintenance and genotyping than by power to detect smaller associations. Inbred plant lines are of course easier to maintain (as seeds) than animal lines, but even so the potential infrastructure costs may be high.

Once the genotypes of the simulated population have been generated, then different subsets can be considered for phenotyping. Valdar et al. (2006) and Klasen et al. (2012) provide guidelines based on simulation for sample sizes required and gains achievable through variation on designs. It is important to note, however, that while these sample sizes may be sufficient to ensure QTL mapping power, other analyses such as high-resolution linkage map construction or epistasis detection may require larger sample sizes. Further simulation may be required depending on the aims for the population. Valdar et al. (2006) compared variations on the first two stages of the collaborative cross (CC) design, and benchmarked against biparental advanced intercross RILs for a trait with 0.5 heritability. They found that a MAGIC population of size 500 could achieve high power to detect single quantitative trait loci (QTL) explaining 5 % of phenotypic variability. Klasen et al. (2012) expanded on these simulation studies by considering a larger range of multiparental designs and heritabilities of 0.5 and 0.8. They found reductions in power when a large number (50–100) of QTL contributed to the trait, but that designs with higher number of parental genomes combined in progeny (such as MAGIC) tended to have increased power.

While simulations are valuable for gaining insight without the expense of a full study, it is worth noting that they of course cannot fully reproduce reality, and at best can consider a limited subset of factors. The simulations above focused on QTL mapping, and to a limited extent, mapping of epistatic interactions, which requires much larger sample sizes to identify effects. Characteristics of genetic data such as introgressions and translocations, which may cause segregation distortion, depression of recombination near centromeres, and other unusual patterns are difficult to study in silico. Hence, with any simulation it is necessary to note the assumptions and the limitations and to be cautious in generalization.

Optimizing resource allocation

While consideration of resource allocation in the design stage is not specific to MAGIC, we include it here due to its relevance for large resource populations. Figure 2 outlines

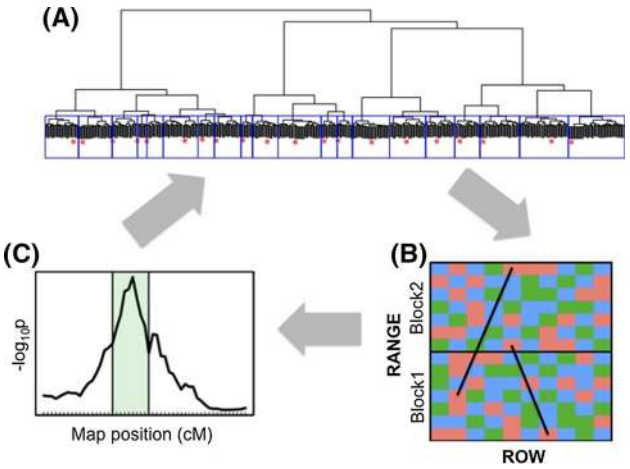

Fig. 2 Efficient phenotyping strategy moving from **a** using SPCLUST to select subset of *lines* for phenotyping by maximizing genetic diversity to **b** planting selected *lines* in a field trial design using partial replication (blocks in *green* and *red*) and compositing (blocks in *red*; *lines* indicating composited samples) to **c** analyzing the data from the field trial accounting for the spatial design. This can become a cycle if QTL support intervals (*green band* in **c**) are then used to select individuals with recombination within the QTL region for future analyses (color figure online)

some general approaches to increase resource allocation efficiency throughout the experiment.

One of the first considerations (once a population has been established) is the selection of lines for phenotyping, as it may not be practical to phenotype the whole population. This is of particular importance for traits which are expensive to measure, or when there are resource constraints, such as when trials require a large amount of land. An alternative to random selection of lines is to attempt to sample representative haplotypes from the population to maximize genetic diversity. Huang et al. (2013) developed an approach based on clustering (SPCLUST) that can efficiently select lines to be phenotyped from a given set of genotyped individuals from a MAGIC population (Fig. 2a). This may be used to: (1) select a general diverse set from a larger population, (2) select based on prior information (e.g., maturity) to subset for important major gene effects, or (3) identify individuals that have breakpoints in a selected region of the genome for fine mapping. After an initial QTL analysis (Fig. 2c), information on candidate regions for QTL may even be input to SPCLUST to prioritize lines for further phenotyping.

Once line selection is complete further efficiency can be gained in crop studies through optimization of resources in experimental design (Fig. 2b). Partially replicated (p-rep) designs were introduced by Cullis et al. (2006), and are commonly used in crops for both research and breeding. Such designs are particularly useful when seed is limited, the trait of interest is expensive to measure and/or where there are multiple phases of experimentation, as in a situation where lines are grown in the field and then further tested in a laboratory (Smith et al. 2006). All of these situations are likely to occur for MAGIC populations due to their size and relevance to a broad range of phenotypes. Combining replicates of lines (composites) from the field for subsequent analysis in the laboratory can provide additional cost benefits (Smith et al. 2015). This approach allows for reduction of the number of samples required in expensive lab testing, while still being able to reduce variability by accounting for spatial effects.

Current status of MAGIC populations in major crops

We summarize the progress of MAGIC studies in three major crops—rice, wheat and chickpea—and provide brief descriptions of other studies in Table 1. While this is not a comprehensive list of all MAGIC populations, it highlights a range of studies with differences in features of the population design and the data collected thus far. These features will impact the analyses possible, and ultimately, the ability to answer specific research questions.

The notation used to present the current status of MAGIC populations was adopted from the published and unpublished datasets of specific crops. F_n denotes the nth filial generation of a cross originally resulting in intercrossed F_1 plants. This differs by a generation from Sn, which denotes the nth generation of selfing of intercrossed F_1 plants. S6:8 is preferred notation for breeders, indicating the selfed generation of the last single plant from which bulked seeds are obtained, e.g., S6, which is equivalent to *F*7.

Rice

Rice (*Oryza sativa* L.) is an inbreeding species with two major ecotypes of cultivated rice: indica and japonica. Indica and japonica rice represent 80 and 20 %, respectively, of the world rice production. Rice is a diploid species with $2n = 24$, and a genome size (of the reference genome Nipponbare) of 473 Mb. Pedigree breeding using biparental populations has been the primary approach in most rice improvement programs. MAGIC is an attractive alternative from both theoretical and practical standpoints. From a theoretical standpoint, MAGIC offers an opportunity to assess the potential of enhanced genetic recombination in trait dissection and synthesis. Although diverse genotypes are regularly introduced into pedigrees, little is known regarding how much genotypic diversity

Table 1 Current status of MAGIC populations in crops

Crop	Design	Progeny	Genotyping	Phenotyping	Analyses	References
Rice	Eight indica cultivars crossed in half-diallel to produce 35 funnels: indica MAGIC (AI0RIL) MAGIC PLUS (AI2RIL) MAGIC PLUS DH (AI2DH)	(May 2014) 1831 S6:8 2206 S6:8 144 S8 48 S4:6 76 DH	GBS ~17,000 SNP	Yield in multiple-environment trials, drought and salinity tolerance, disease resistance, grain quality	QTL, GWAS	Bandillo et al. (2013)
Rice	Eight japonica cultivars crossed in half-diallel to produce 35 funnels: japonica MAGIC (AI0RIL, in progress)	(May 2014) 400 S4:6	–	–	–	Bandillo et al. (2013)
Rice	Eight indica + eight japonica cultivars, each crossed in half-diallel with 35 funnels, then intercrossed: MAGIC GLOBAL (AI0RIL)	(May 2014) 1402 S7	–	–	–	Bandillo et al. (2013)
Wheat	Four spring wheat cultivars, 1 funnel, AI0RIL	1579 F6	826 DArTs, 283 SNPs and 53 SSRs; 9K SNP array; 90K SNP array	Plant height, hectolitre weight, coleoptile length	LMC, QTL	Huang et al. (2012); Cavanagh et al. (2013); Rebetzke et al. (2014)
Wheat	Eight spring wheat cultivars crossed in half-diallel with 315 funnels: AI0RIL AI2RIL AI3RIL	2099 F6 367 F6 473 F6	9K SNP array; 90K SNP array; GBS	–	LMC, QTL	CSIRO (unpublished)
Wheat	Eight winter wheat cultivars, crossed in half-diallel with 315 funnels, AI0RIL	1091 F7	90K SNP array	Yield, flowering time, plant height, yellow rust, fusarium, mildew, awn presence	QTL, GWAS	Mackay et al. (2014); Scutari et al. (2014)
Wheat	16 spring wheat (in progress), 125 funnels, AI0RIL	600 RILs	–	–	–	CSIRO (unpublished)
Wheat	16 winter wheat, crossed in half-diallel with 15 funnels, AI0RIL	800 RILs	–	–	–	NIAB (unpublished)
Wheat	60 parents randomly intercrossed, AI12RIL	1000 F3	9K SNP array	Heading date	LDA, GWAS	Thépot et al. (2015)
Chickpea	Eight *desi* (complete)	1000 F6	Re-sequencing ~1000 lines in progress	–	–	Gaur et al. (2012)
Chickpea	Eight *kabuli* (in progress)	–	–	–	–	ICRISAT (unpublished)
Pigeonpea	Eight (in progress), 7 funnels	At S1 stage	–	–	–	ICRISAT (unpublished)
Peanut	Eight (in progress), 14 funnels	At S1 stage	–	–	–	ICRISAT (unpublished)
Maize	Eight parents crossed in half-diallel with 35 funnels, AI0RIL	>1300 F6	50K SNP chip; GBS	Plant height, ear height, yield, flowering time	–	Pea et al. (2013)

Table 1 continued

Crop	Design	Progeny	Genotyping	Phenotyping	Analyses	References
Maize	Sixteen historical lines representing heterotic groups used for hybrid production in temperate regions	543 DH	MaizeSNP50 beadchip; 400K SNP	Multiple environments since 2011	–	Buet al. (2013)
Barley	Seven landraces, one modern variety, 1 funnel, AI0 DH	~5000 AI0 DH; 534 genotyped and phenotyped	9K SNP chip	Flowering time	LMC, QTL, GWAS	Sannemann et al. (2015)
Oat	Eight spring oat crossed in half-diallel with 42 funnels AI0RIL	600 S6	6K SNP chip			IBERS, Aberystwyth University (unpublished)
Durum wheat	Four cultivars, 1 funnel, AI0RIL	338 F7	90K SNP chip	Days to heading and maturity, plant height, grain yield	LMC, QTL, GWAS	University of Bologna (unpublished)
Tomato	Four normal tomato, four cherry tomato, 1 funnel, AI0RIL	397 S3	1536 custom SNP chip	Fruit weight	LMC, QTL, GWAS	Pascual et al. (2015)

F_n denotes the *n*th filial generation of a cross originally resulting in intercrossed F_1 plants. This differs by a generation from Sn, which denotes the *n*th generation of selfing of intercrossed F_1 plants. S6:8 is preferred notation for breeders, indicating the selfed generation of the last single plant from which bulked seeds are obtained, e.g., S6, which is equivalent to F7

AIx advanced intercrossing for × generations, *RIL* inbreeding to create recombinant inbred lines, *DH* doubled haploids, *GBS* genotyping-by-sequencing, *LMC* linkage map construction, *QTL* QTL mapping, *GWAS* genome wide association mapping, *LDA* linkage disequilibrium analysis

created by recombination has been captured, particularly for achieving genetic gain in complex traits such as yield. From a practical standpoint, a compact genetic resource with moderate population size and a concentration of high-value traits is particularly valuable as a pre-breeding gene pool.

Four rice MAGIC populations are currently at different stages of development in the Philippines (Bandillo et al. 2013). The furthest progressed are the indica MAGIC and MAGIC PLUS populations, which are both derived from eight indica parents and intercrossed for zero or two generations, respectively. The indica MAGIC population consists of 60 plants from each of 35 funnels, generating a population of 2100 RILs. The MAGIC PLUS population has the same first mixing stage, but produces 12 lines from each of 175 crosses in the second AI stage, generating the same-sized population. The eight parents, 200 S4 indica MAGIC RILs, and 190 MAGIC PLUS S4 RILs have been genotyped using genotyping-by-sequencing (Elshire et al. 2011). The other two populations incorporate japonica lines, which is the subtype to which the reference sequence (cultivar Nipponbare) belongs. The japonica MAGIC follows the same population design as the indica MAGIC, while the Global 16-parent MAGIC intercrosses eight-way lines from the indica MAGIC and japonica MAGIC prior to inbreeding via production of RILs.

The rice MAGIC populations were deliberately developed to serve breeding applications. Rice MAGIC lines are presently being extensively phenotyped, presenting new challenges for dissection of complex traits such as yield, drought tolerance, and quantitative disease resistance. Approximately 17,000 single nucleotide polymorphism (SNP) markers have been used in genome wide association mapping for multiple traits, including: blast and bacterial blight resistance, salinity and submergence tolerance, and grain quality. Over 1000 lines have been extracted from the different MAGIC populations by IRRI breeders and are being tested in market environments for use in breeding programs. High-yielding lines with favorable agronomic traits have been identified. Rice breeders consider rice MAGIC lines as good pre-breeding material with a package of favorable traits.

The success so far has led to the development at IRRI of additional MAGIC populations focused on the traits biotic stress tolerance and heat tolerance. Lines have also been distributed for evaluation in multiple countries in Southeast Asia (Indonesia, Vietnam, Laos, Myanmar, Taiwan, Philippines) and Africa (Tanzania, Senegal). Yield measurements from these multi-environment trials exhibit transgressive segregation. These data offer ideal materials for studying gene × gene and gene × environment interactions.

Wheat

Bread wheat (*Triticum aestivum* L., $2n = 6\times = 42$, genome size 17 Gb) is one of the most important crops globally, with a worldwide production of over 710 million metric tons. The complexity of its genome structure requires novel approaches for its genetic dissection, leading to great uptake of MAGIC among wheat researchers. The first MAGIC wheat population was grown in Australia (Huang et al. 2012) and inter-mated four spring-type Australian cultivar parents. Subsequently, eight-parent and 16-parent populations have been created both in Australia (spring wheat) and the United Kingdom (winter wheat; Mackay et al. 2014). The basic designs for the eight-parent populations are similar to those for the indica rice population, namely, mixing all eight founders through multiple funnels with no intercrossing to create RILs. However, the wheat designs are more ambitious, realizing lines from many more funnels. Subsets of both eight-parent wheat populations also created lines with two to three generations of intercrossing.

The wheat populations are the furthest progressed in data collection, with over 1000 lines genotyped in winter wheat, and nearly 4000 in the two spring wheat MAGIC populations. This has motivated development of numerous methods for analysis, both in the realm of linkage map construction (Huang and George 2011; Ahfock et al. 2014), and marker–trait association mapping (Verbyla et al. 2014a, b; Scutari et al. 2014). Extensive data have been collected and analyzed on traits including yield, disease resistance, plant height, flowering time, and coleoptile length (Huang et al. 2012; Rebetzke et al. 2014; Mackay et al. 2014; Scutari et al. 2014). Comparison of results in these populations with those from biparental populations is discussed in later sections on genetic analysis.

Chickpeas

Chickpea (*Cicer arietinum* L., $2n = 2\times = 16$ and genome size is ~738-Mb), the second largest consumed pulse crop of the world, is grown in over 50 countries and traded across 140 countries. Similar to rice, there are two major ecotypes: *desi*, which have a small seed and a dark, thick seed coat, and *kabuli*, which are larger, lighter colored, with a smooth coat. However, the genetic base of the elite genepool is very narrow and poses a serious constraint both for breeding and for mapping traits of interest. To enhance the genetic base and identify marker–trait associations for target traits, efforts have been made to develop the first set of the *desi* MAGIC population. As with both the wheat and rice eight-way populations, a total of eight elite and diverse founder parents were selected and crossed in half-diallel mating design to develop 28 two-way F_1s. These lines were subsequently combined to produce 14 four-way intercrosses and finally, seven funnels. Selfing of lines from the seven funnels through single seed descent resulted in the development of ~1000 F_6 MAGIC lines.

Currently, the eight parents have been sequenced at 8–10× coverage, and ~1000 MAGIC lines are being resequenced at lower coverage (2–3×). These data will be used to classify the MAGIC lines into different groups based on SNPs and haplotypes. Subsequently, a set of 200–500 lines possessing non-redundant haplotypes will be identified and used for extensive phenotyping for targeted traits in multiple environments. Genotyping data and phenotyping data collected on the set of MAGIC lines will be analyzed for establishing marker–trait associations via genome wide association studies (GWAS) for targeted traits. In addition, those MAGIC lines well characterized at both the molecular and phenotypic level will be an ideal resource for deploying in chickpea breeding programs.

While there has been slower uptake of the MAGIC design in other crops, a four-parent durum wheat MAGIC population has been produced (S. Milner, pers. Comm.), and eight-parent MAGIC resources are in production for maize (M. dell'Acqua, pers. Comm.), barley (Sannemann et al. 2015), pigeonpea (C. Sameerkumar, pers. Comm.), peanut (P. Janila, pers. Comm.), and tomato (Pascual et al. 2015). Brief summaries of these populations are listed in Table 1.

Data management

The suitability of MAGIC populations as focal points for communities of researchers and foundations for in-depth system biology analysis necessitates a shift in thinking regarding the collection and maintenance of associated data. Genetic, phenotypic, environmental, and 'omics data may all be collected on individuals with the same genetic makeup, with many of these variables changing value over time. The integration of these data in analysis requires efficient storage and distribution of the data.

Management of large genetic resource populations

In most cases, these collections of genotypic and phenotypic data are/should be shared across multiple groups of researchers. The mouse Collaborative Cross Consortium has had the most experience in this area, as even the development of the population was split across multiple locations (Collaborative Cross Consortium 2012). Durrant et al. (2012) review existing tools and evaluate the need to further integrate resources for both bioinformatic analysis and data management in mice. Researchers can additionally

look at human studies to learn from the lessons of large-scale multicentre clinical trials in handling big data of this type (e.g., Das et al. 2011).

Big data management is not a new problem (Schmitt and Burchinal 2011); indeed, in 2011 Science had a special online collection dealing with data highlighting the ubiquity of the problem across diverse scientific disciplines. However, it is an emerging problem for crop studies, which in the past were rarely of this scale. There is still a pressing need for data infrastructure to be recognized as a critical part of studies of this size, and for well-developed systems to become the norm rather than the exception. We provide details of some systems available for breeding management and analysis in crops (Table 2).

High-throughput genotyping

Efficient and consistent handling of genetic data is a priority for data management in MAGIC studies. While in theory, a resource population needs only be genotyped once if done at sufficiently high density and quality, in practice most populations have been genotyped more than once as genotyping technologies evolve. In wheat, it has progressed from Diversity Arrays Technology (DArT) markers and simple sequence repeats (SSRs) (Huang et al. 2012) to a 9K SNP chip (Cavanagh et al. 2013), and finally a 90K SNP chip (Wang et al. 2014). As genotyping-by-sequencing technologies develop, the relative low cost of this approach will most likely result in great uptake as in the case of rice and chickpea.

Issues can arise in maintaining coherency among these data, primarily due to changes over time. Individuals may be genotyped at different generations of inbreeding, resulting not only in differing levels of heterozygosity, but also differences in genome structure at subsequent generations due to differential fixing of alleles at formerly heterozygous loci. As long as all individuals are substantially progressed through inbreeding prior to genotyping, this should be a minor issue, but it may result in subtle differences in analysis. Different individuals may be genotyped on different platforms, resulting in systematic missing data. Different platforms may have different biases and quality of data, such as the propensity of GBS to have high levels of missing data (Elshire et al. 2011). Even for a single platform, reductions in cost over time may result in higher coverage for individuals genotyped a few months later than others. Many of these issues will require development of methods for genotype imputation (http://mus.well.ox.ac.uk/19genomes/magic.html; Wang et al. 2012; Huang et al. 2014) and analysis approaches which accommodate potential confounding factors.

Table 2 Data management platforms for plants and crops

Platform	Availability	Functionality	Website
Agrobase	Proprietary	Windows-based relational software system; data management, experimental design, statistical analysis with interface to GenStat	http://www.agronomix.com/
iPlant	International collaborative, free of charge to collaborators intending to make data public	Cloud-based data store; high-performance computing with virtual servers; bioinformatics pipelines; image analysis	http://www.iplantcollaborative.org/
Integrated Breeding Platform (IBP)	Non-profit organization; Low to no cost, dependent on organization and location	Wide range of products from breeding management system to analysis, either desktop or cloud based; partnered with iPlant	https://www.integratedbreeding.net/
Phenome networks	Software as a service provider; currently beta and free to academic users	Web-based network to host, manage, analyze, and share phenotypic and genotypic studies; breeding management software	http://phenome-networks.com/
KDDart	Available to DArT clients; Currently used by early adopter DArT clients and collaborators	Integration of genetic, phenotypic, and environmental data; applications to automatically collect data, perform analyses, and interface with sensors	http://www.diversityarrays.com/kddart

High-throughput phenotyping

The potential phenotypic diversity of the MAGIC populations makes them ideally suited for high-throughput phenotyping endeavors. High-throughput phenotyping can be interpreted in two senses for these populations. The first occurs at a molecular level, with different 'omics' platforms producing thousands of traits representing gene, protein or metabolite expression. The second occurs at the individual plant level, however, capitalizing on the wide range of traits segregating in MAGIC, and the ability to phenotype traits for the same inbred line across different environments, years, and conditions. In both cases, implementing and profiting from the high information content of the data rely on novel technologies.

Measurement of different 'omics' traits has thus far been limited to MAGIC populations in model organisms. In pre-Collaborative Cross mice (not fully inbred), Aylor et al. (2011) and Bottomly et al. (2012) performed expression QTL (eQTL) studies to investigate associations with gene expression. Even in a relatively limited sample of 220 lines, Aylor et al. (2011) detect abundant eQTLs, more than double the number reported in other mouse studies. In Drosophila, King et al. (2014) identified nearly 8000 eQTL, predominantly *cis*, but with a number of *trans*-eQTL hotspots, indicating eQTL regulating the expression of several other genes. The success of these studies is promising for 'omics' analysis in crops.

A far more typical situation in crops, however, arises in attempting to phenotype a large number of lines in glasshouse or field trials. This can be time-consuming and expensive, and ultimately, limit the number of lines phenotyped and/or traits investigated. Automation of phenotyping is an active area of research, with several reviews published recently describing advances in sensors, robotics and imaging which may revolutionize the collection of data for large field trials (Montes et al. 2007; Furbank and Tester 2011; Araus and Cairns 2014). These approaches will be crucial for fully capitalizing on the strengths of MAGIC, not just in field trials, but also for controlled environment experiments.

Genetic analysis

Analysis of MAGIC populations has many similarities to that of biparental populations, but it must accommodate the unique features of the design such as multiple founder alleles. This makes it impossible or at best inefficient to apply previously developed methods and has stimulated the development of a number of MAGIC-targeted software tools (Table 3). By incorporating the known family structure and the ability to differentiate between founders, analysis approaches can gain in power, precision and depth of interpretation for these populations.

Linkage map construction

The large number of polymorphic markers across all founders and accumulation of recombination events through many generations of the MAGIC pedigree can be used to achieve dense and high-resolution mapping of the genome. The first linkage map from a MAGIC population was constructed in wheat (Huang et al. 2012), which due to its large genome size (17 GB) and hexaploid nature, does not yet have a reference sequence. When maps from six biparental wheat populations were combined with a four-parent MAGIC map to create a consensus map of markers from

Table 3 Software tools designed for the simulation and analysis of multi-parent populations (MPP)

Software package	Applicability	Functionality	Availability	References
HAPPY	General MPP	QTL analysis; permutation	http://mus.well.ox.ac.uk/magic/	Mott et al. (2000)
R/qtl	4-way, 8-way, 16-way MAGIC	Simulation; map construction; QTL analysis; imputation	CRAN	Broman et al. (2003)
R/ricalc	MAGIC by selfing, sib-mating	Simulation; probability calculation	https://github.com/kbroman/ricalc	Broman (2005)
Genome_scan	General MPP; full sequence	QTL analysis; permutation	http://mus.well.ox.ac.uk/19genomes/magic.html	
R/mpMap	4-way, 8-way MAGIC by selfing	Simulation; map construction; QTL analysis; imputation	https://github.com/behuang/mpMap	Huang and George (2011)
R/spclust	NAM, 4-way, 8-way MAGIC by selfing	Selective phenotyping	https://github.com/behuang/spclust	Huang et al. (2013)
R/mpwgaim	4-way, 8-way MAGIC by selfing	QTL analysis	Contact authors	Verbyla et al. (2014a)
AlphaMPSim	General MPP	Simulation	https://sites.google.com/site/hickey-john/workstuff/alphampsim	Hickey et al. (2014)

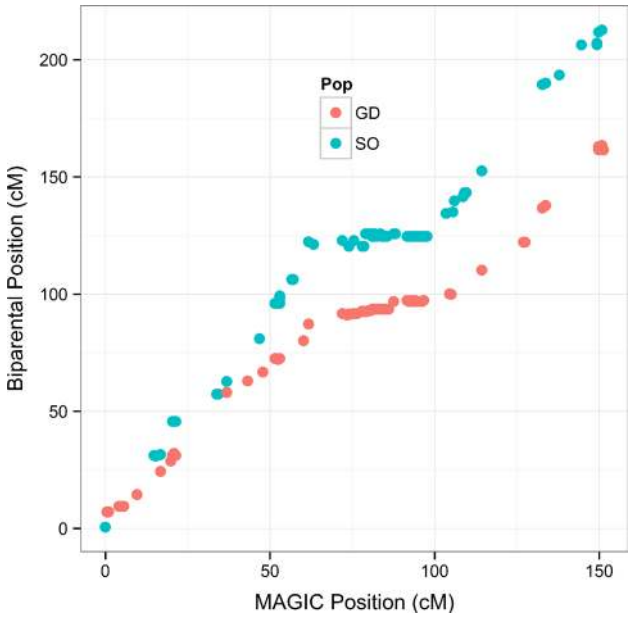

Fig. 3 Comparison of wheat 9K SNP maps of Chr 3A in MAGIC four-way population with those from two biparental populations—Gladius × Drysdale (GD, orange) and Synthetic × Opata (SO, *green*). *Markers* in the middle of the chromosome display little recombination in the biparental populations, but are spread over more than 20 cM in the MAGIC map (color figure online)

a 9K SNP chip, 7504 polymorphic loci in total were positioned (Cavanagh et al. 2013). Of these, 3931 were mapped in the four-parent MAGIC population, while the median number mapped in biparental populations was 2670. Further, these markers tended to represent unique recombination bins more frequently in MAGIC than in the biparental populations. The higher levels of recombination in the MAGIC population can be seen most clearly in the region around centromeres, where there is very little separation between markers in biparental populations (Fig. 3).

While it can be difficult to validate map order derived from these populations, Ahfock et al. (2014) suggest methods to assess map uncertainty which may be particularly relevant for high-density maps. Further, they suggest guidelines for sample sizes and marker types in multiparental populations to ensure low uncertainty in map positions. In particular, for four-parent MAGIC population sizes of 500 lines, most markers spaced at least 2 cM apart can be confidently ordered, while for larger populations of 1000 lines this improves to 1 cM resolution. As resources improve, availability of a physical map or reference sequence will provide a standard against which to validate the genetic map; conversely, the high-density genetic maps achievable from MAGIC populations may be used to verify and anchor sequence assemblies.

The advantages of MAGIC populations for map construction have accompanying challenges, however. In contrast to biparental populations, recombination events cannot be directly observed due to the use of biallelic markers. Thus, the estimation of recombination fractions is more computationally demanding, particularly for high-density mapping. Further, the pattern of alleles among the founders has a major impact on the accuracy of mapping, with the recombination fractions between certain alleles not being identifiable (Ahfock et al. 2014). This may require imputation of recombination fractions or removal of certain markers prior to map construction.

Haplotype mosaic reconstruction

Once a high-density map or reference sequence has been established, it serves as a foundation for investigation of genome structure relevant both within the MAGIC population and to the species in general. In most studies, genetic data are represented as marker scores resulting from genotype calling, whether these are biallelic SNPs, multiallelic SSR, or even bases at different sequence positions. In MAGIC populations, these serve as surrogates for the underlying (unobserved) alleles inherited from each founder. Combining our knowledge of the pedigree structure with the observed data allows us to probabilistically reconstruct the haplotype mosaics which represent the mixing of the founder genomes to produce inbred lines (Fig. 4).

The different stages of this haplotype reconstruction process have led to three representations of genetic data in MAGIC populations: marker scores, founder probabilities, and the mosaics themselves. Each can be used to investigate genomic structure and as input to QTL analysis. To estimate the founder probabilities, Hidden Markov Model (HMM) methods are typically used, which may depend on the pedigree structure (Broman et al. 2003) or just the observed data (Mott et al. 2000). The mosaics are one step further removed from the marker data, essentially imputing multiallelic markers from the probabilities. The drawback is that this approach does not account for the uncertainty associated with the estimation process; further, in regions of the genome where founder genotypes are very similar or marker density is low, both of which result in lower genomic information content, it may not be able to impute alleles with high certainty. However, if estimated with low error, these values should be closest to the true underlying genotypes, and can be used to identify recombination breakpoints in individual lines.

Haplotype mosaic reconstruction from high-density genotype data allows positions of recombination breakpoints to be estimated with high accuracy. This supports identification of recombination hotspots and QTL for recombination events. While this is also feasible in biparental RIL populations (Esch et al. 2007), the greater

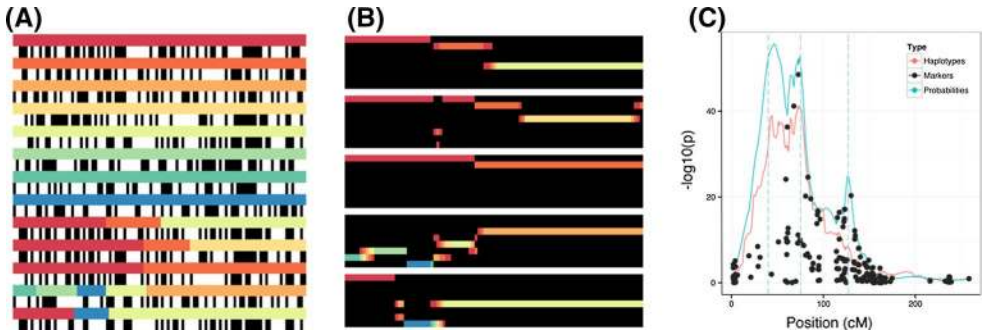

Fig. 4 Comparison of different representations of simulated genetic data in MAGIC populations. **a** Example haplotype mosaics (*colored lines*) and SNP data (*black* and *white lines*) for founders (first eight) and five RILs (last five); **b** construction of RIL haplotype mosaics by finding best path through estimated founder probabilities. *Colored segments* indicate most likely founders in that region; **c** example of simple interval mapping based on probabilities (*green line*); haplotypes (*red line*); and biallelic SNPs (points); *dashed blue lines* denote true locations of QTL (color figure online)

resolution offered by MAGIC may provide increased insight into genome structure. In theory, the different variations on MAGIC designs will produce additional benefits for investigating genome structure. In rice, previous studies using SSR and restriction fragment length polymorphism (RFLP) markers (Harushima et al. 1998) and low-depth sequencing data of biparental RILs suggested that each meiosis produced approximately 33 recombinations per genome (Huang et al. 2009). Yamamoto et al. (2014) used simulation studies to estimate recombination frequency in rice MAGIC populations. Their results suggest approximately 160 genome segments per individual with a mean segment size of 20 cM in a multi-parent population with no cycles of recurrent crossing. However, comparison of the MAGIC Indica and MAGIC PLUS, which underwent two additional generations of intercrossing, did not confirm the theoretical predictions of the simulation study. More samples and more complete data may be required to determine the recombination breakpoints more accurately.

QTL mapping approaches

The first step in mapping the relationship between phenotype and genotype and detecting gene-trait associations is often identifying QTL. The development of methods to utilize the additional information available in MAGIC populations has been essential to maximize the usefulness of such resources. Methods can differ in a number of ways. Inputs may be marker scores or founder probabilities; search strategies may involve a genome scan, or modeling of QTL while simultaneously accounting for all other markers genotyped. Statistical approaches can be frequentist or Bayesian, differ in type of model and the number of stages used (1-stage vs. 2-stage), and in how the QTL effects are modeled (fixed vs. random). Figure 5 illustrates the relationship and differences between various methods applied to MAGIC populations.

The most common approach to QTL analysis in MAGIC populations is to use a genome scan such as interval mapping, testing each marker or interval separately for association with the trait of interest. Xu (1996) demonstrated the first QTL analysis for a four-way cross using markers in an interval mapping approach based on the regression method of Haley and Knott (1992). However, this approach failed in a mouse heterogeneous stock population (Mott et al. 2000), leading to development of HAPPY, an interval mapping approach method based on founder probabilities. This and a similar approach were applied to the *A. thaliana* MAGIC (Kover et al. 2009) and Drosophila populations (King et al. 2012b).

In spite of the difficulty encountered by Mott et al. (2000) using markers, several studies have employed an association mapping approach (Bandillo et al. 2013; Mackay et al. 2014), using existing GWAS software such as TASSEL (Bradbury et al. 2007). The benefit of this approach is that the test is simple and computationally straightforward, requiring fewer degrees of freedom than other methods. Such methods typically agree well with those based on founder probabilities for large QTL, but may differ in performance for those with smaller effects. An example comparing the performance of interval mapping based on marker scores, founder probabilities, and haplotype mosaics is shown in Fig. 4c. In general, differences between the methods can be attributed to the ability to accurately estimate probabilities, coupled with the true number of allelic effects among the founders.

Various compromises between marker scores and founder probabilities have been proposed to provide a flexible and parsimonious model. Yalcin et al. (2005) suggest testing based on imputing the genomes of the mapping population from the founder probabilities and founder

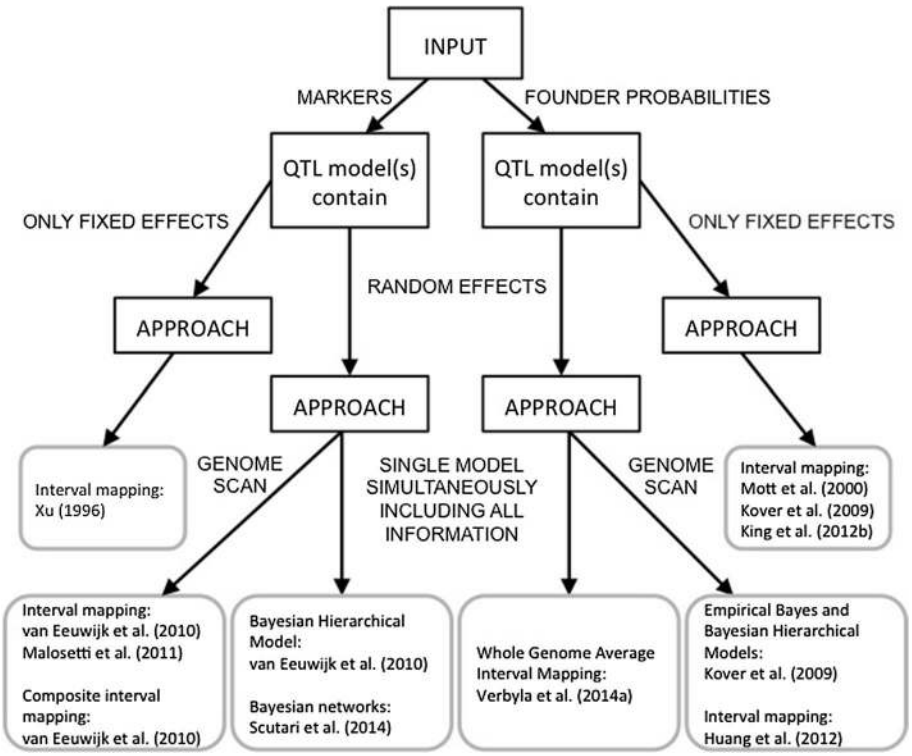

Fig. 5 Deconstruction of existing QTL mapping approaches for MAGIC populations based on common analysis features

sequence information. This retains a biallelic model, but uses haplotype information to inform the genotypes. A more general approach is based on linkage disequilibrium and linkage analysis (LDLA) mapping, originally proposed by Meuwissen and Goddard (2001). LDLA models local similarities using haplotypes, but does not require all founders to have separate effects. This can increase power when there are many founders or founders are genetically similar. Such models have had greater uptake in linked biparental populations than in MAGIC studies (e.g., Giraud et al. (2014)), though their use has been investigated in a durum wheat MAGIC (S. Milner, pers. comm.), using the ClustHaplo (Leroux et al. 2014) software to create haplotypes. While no simulations of their efficacy have been performed yet in MAGIC, Bardol et al. (2013) compared a variety of models ranging from biallelic to individual founder effects in data from a maize breeding program and concluded that the different models were complementary depending on the QTL and trait under consideration.

Interval mapping approaches often proceed in two stages, modeling phenotypic variation prior to genetic variation to reduce the computational burden. However, one-stage methods allow greater model flexibility and can avoid complexities associated with correctly accounting for estimation error across two stages (Mohring and Piepho 2009). The approach described in Huang and George (2011) expands the basic HAPPY model to a linear mixed model framework to simultaneously model genetic and spatial or environmental variation. This was applied in Huang et al. (2012) for the analysis of plant height and hectoliter weight to demonstrate that known genes could be detected with increased precision (confidence interval half as wide) relative to biparental populations. Kover et al. (2009) compared HAPPY to an empirical Bayes method combining founder probabilities with random effects to model population structure, but found little difference in results. A similar approach based on markers was used by Malosetti et al. (2011) for a three-way cross in barley.

Recognition that interval mapping approaches can lead to bias in QTL mapping led to the development of more sophisticated approaches such as composite interval mapping (CIM; Zeng 1994; Jansen 1994) and multiple interval mapping (Kao et al. 1999). By including other covariates to account for background variation, these approaches more accurately model the magnitude and location of QTL. While developed for biparental populations, they can be relatively easily extended to MAGIC populations (Huang and George 2011; van Eeuwijk et al. 2010).

Expanding upon the idea of incorporating covariates, Verbyla et al. (2007) proposed to accommodate all genotype information in a single model with whole genome average interval mapping (WGAIM). The original approach for biparental populations was demonstrated to be more powerful than CIM, and has been extended for use in MAGIC populations (Verbyla et al. 2014a) and applied in wheat by Rebetzke et al. (2014). The approach

(multi-parent WGAIM or MPWGAIM) adds QTL to the model through forward selection, while modeling effects from population structure and any non-genetic effects, such as spatial variation, in a linear mixed model framework. A dimension reduction technique is implemented to reduce model complexity.

Bayesian methods offer a flexible alternative to mixed models, but are typically computationally demanding, requiring many Markov chain Monte Carlo iterations to reach convergence. Kover et al. (2009) use a hierarchical Bayes method to perform a genome scan, explained more fully in Durrant and Mott (2010). As the choice of priors for this method results in complete factorization of the likelihood, no Monte Carlo Markov chains are required, resulting in a very fast Bayesian approach. van Eeuwijk et al. (2010) present a whole genome method based on the work of Bink et al. (2008) that includes all markers in a single hierarchical Bayesian model where the number of QTL is allowed to change through the use of a reversible jump algorithm (Green 1995).

Determination of significance thresholds is an important aspect of QTL analysis. The genome scan methods typically use resampling techniques (permutation tests or the bootstrap). In contrast, the Bayesian approaches use Bayes factors (Kass 1993; Kass and Raftery 1995) to assess the support for different QTL models, and MPWGAIM uses a likelihood ratio test of significance at a set type I error rate to determine when to cease forward selection.

The computational demands of any approach are dependent on the complexity of the statistical models implemented. Approaches using linear models to perform genome scans can often be parallelized easily and can, therefore, be more computationally efficient than those relying on complex linear mixed models. For example, software developed for the Arabidopsis MAGIC population (genome_scan, http://mus.well.ox.ac.uk/19genomes/magic.html) can analyze full sequence data very quickly, fast enough that permutations are a viable method of computing genome wide significance. Reducing the complexity of the final model using multiple stages is often an option. However, the computational gain must be balanced against potential loss of statistical efficiency; with certain experimental designs (e.g., p-rep), this may be a reason to consider more complex models.

Future prospects

MAGIC populations have proven their worth for standard genetic analysis—linkage map construction, linkage analysis, and association analysis—but the real test of their longevity will be whether these results can be extended beyond preliminary identification of interesting genomic regions. In this section, we take a look forward and discuss some of the areas we think will provide a valuable contribution to our understanding of complex traits into the future.

Multivariate analysis

In crop studies, material is often grown in multiple environments and multiple traits are measured; sometimes a trait is measured at several time points. The analysis of such situations is an area of particular promise for inbred populations in general and MAGIC specifically, as the inbred lines can be easily replicated across environments. Further, MAGIC populations have been developed to capitalize on phenotypic diversity, so a multitude of traits segregate in each population. We call these complex situations *multivariate*, and the methods required for their analysis are correspondingly complex, in biparental populations as well as multiparental. The complexity arises from both the genetic and non-genetic components of the model, and the analysis is expensive both in time and computing resources. As fitting methods are iterative, there can be issues with convergence. Automating such analyses is very difficult.

Multivariate analyses build on underlying correlation, thereby providing more powerful analysis. For multi-environment trials, this leads to an understanding of genotype (or QTL) by environment interaction. For multi-trait analyses, this furthers our understanding of pleiotropic QTL. Both types of information can provide valuable input to a breeding program.

Multivariate QTL analysis for biparental populations has a modest literature, much of which has been previously outlined (Verbyla and Cullis 2012; Malosetti et al. 2013). For MAGIC, methods are only beginning to become available; Verbyla et al. (2014b) extend MPWGAIM (Verbyla et al. 2014a) to the multivariate case, while Scutari et al. (2014) take an alternative approach using Bayesian networks. In addition, Malosetti et al. (2013) mention that their multivariate models for biparental populations can be easily extended to multiparental populations. While both MPWGAIM and Bayesian networks have been applied to wheat MAGIC data, a direct comparison has not yet been made. This is in part due to differences in focus; while the Bayesian network approach can be used to identify associations, it is primarily intended for prediction. For all methods, analyses can be time-consuming and require large computing resources due to the complexity of models being fitted. There is a need to investigate two-stage analyses for this purpose. Although they may be less statistically efficient, they are likely to be less computationally demanding.

Epistasis detection

One of the main benefits of the MAGIC populations is the creation of new combinations of alleles through generations of mixing founder genomes together. However, the detection of epistatic interactions in MAGIC has been little explored to date. One of the few in-depth studies of epistasis was performed in the context of two-locus segregation distortion (Corbett-Detig et al. 2013), and considered biallelic interactions in the Arabidopsis MAGIC as well as other multiparental populations. A few studies (e.g., Huang et al. 2011, 2012) have considered interactions between QTL detected with main effects. While such a strategy may be successful in identifying a few instances of epistasis, a full genome wide scan may be necessary to detect epistasis between loci without main effects. Indeed, Scutari et al. (2014) indicate that a disadvantage of their Bayesian network approach is its limited ability to capture smaller epistatic effects, although they should be able to capture those where one of the SNPs has a main effect.

There are three primary issues with escalating the search for epistasis in these populations. The first is computational—as the number of genetic markers available for most populations increases from the thousands to the hundreds of thousands, the number of pairwise interactions to test increases by four orders of magnitude. However, exhaustive search methods have become feasible in human association studies with the application of parallel computing, either with Graphics Processing Units (Hemani et al. 2011) or computing clusters (Gyenesei et al. 2012). While these can only be applied to test for interactions between observed marker scores, extension to accommodate multiple founders will have great utility in MAGIC populations. The second issue, however, actually stems from the multiplicity of founder alleles. Testing for differences between eight parental effects, while at times less powerful than comparing two alleles, is still a low-dimensional endeavor. Testing for differences between 64 pairwise combinations, however, may not be feasible except in large samples. Hence, the third issue is power; depending on how the first two issues are addressed, the sample size required to detect epistatic interactions may be much larger than that required to detect main effects of similar magnitude. Determining the best approach to modeling epistatic interactions in MAGIC requires further study through simulations prior to application in real data.

Multi-parent advanced generation recurrent selection (MAGReS)

MAGIC populations and their derivatives additionally offer the opportunity to develop genotypes with combinations of superior alleles from diverse backgrounds, and thereby directly improve a breeding program. One option is to apply genomic selection to the RIL progeny for forward selection via genotype only. We propose here the alternative of *multi-parent advanced generation recurrent selection (MAGReS)*, for which the breeding scheme is presented in Fig. 6. This approach combines MAGIC for the development of recombinant lines, with marker-assisted recurrent selection (MARS) for identification of superior lines for intercrossing, to develop lines possessing elite alleles from diverse parents. The MAGReS lines will be highly diverse and can be used as donors for different breeding lines, leading to direct release of superior lines as new varieties for commercial cultivation.

MAGReS initially follows the same stages of population development as MAGIC, with the further constraints upon founder selection of breeder-relevant alleles for a trait of interest, and high molecular diversity. Once MAGIC lines have been developed, QTL mapping approaches can be employed to identify associations with target traits. These associations in turn are used to select plants possessing the maximum number of positive/superior alleles. Two to three generations of intercrossing between the selected plants are then used to assemble all targeted superior alleles from the diverse founder parents in a common background. It is important to bear in mind that all traits of interest should ideally be considered concurrently, as this process has the potential to break up some existing allelic combinations for traits which have not been targeted, which could result in lower or unacceptable performance for these traits. Finally, newly developed breeding lines possessing all the target alleles can be selfed to identify inbred lines homozygous for the target alleles. Such lines form a direct pipeline to commercial cultivation through phenotyping for the target traits in multilocation trials and selection of improved lines showing positive response.

Multi-parent populations such as MAGIC provide a unique opportunity to explore genetic diversity, better define genomic intervals involved in complex traits, model complex environmental interactions and better predict allelic effects in diverse backgrounds. In species where a genome reference sequence is not available, these populations also provide a valuable resource for generating high-density linkage maps that may be used for anchoring the meiotic map with the physical map.

However, MAGIC designs are not the only type of multiparental populations, and different designs may be more or less suitable depending on the crop, its genome structure, and resources available (Stich 2009). Nested Association Mapping (NAM) designs were proposed by Yu et al. (2008) and have had uptake in several crops including maize (McMullen et al. 2009; Giraud et al. 2014), barley (Schnaithmann et al. 2014) and sorghum (Mace et al. 2013). The strategy of using related biparental populations

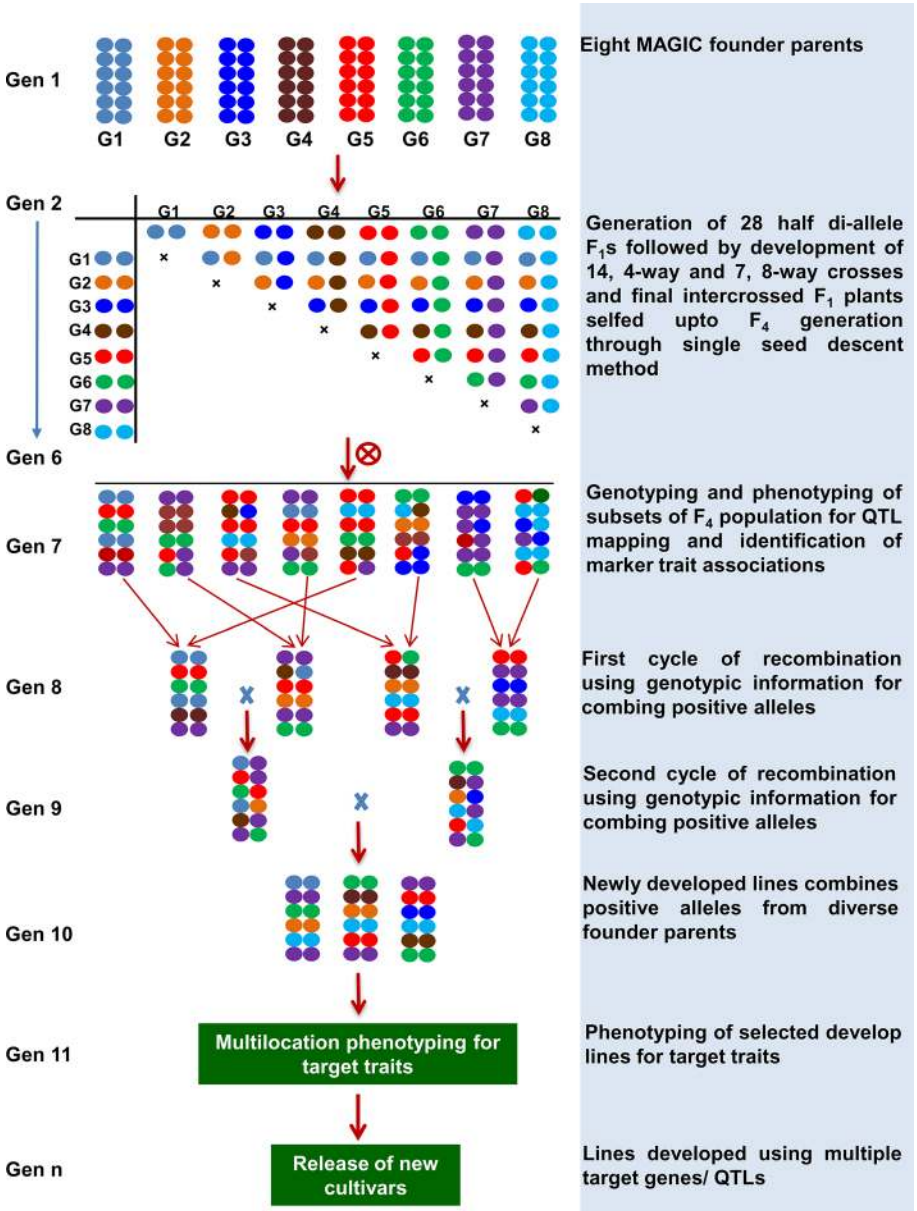

Fig. 6 Multi-parent advanced generation recurrent selection (MAGReS) approach for development of new breeding lines. Development of MAGIC lines is followed by QTL analysis from which markers associated with the trait of interest can be identified. MAGIC lines possessing large numbers of desired alleles are selected and combined for 2–3 additional cycles of recombination through marker-assisted recurrent selection (MARS), leading to the development of superior breeding lines. G1, G2, G3, G4, G5, G6, G7 and G8 represent the 8 diverse founder genotypes; *Gen* generation (color figure online)

requires less time to develop than MAGIC populations. Further, the choice of founders is less crucial since the populations can be post-modified to either include lines derived from additional founder crosses, or exclude those derived from an undesirable founder. However, fewer novel allelic combinations will be generated in these populations, as any specific individual carries alleles from only two founders rather than all of them.

Multiparental populations of all types are still in their infancy, and their value will be judged through their ability to deliver solutions and understanding of the genetic determinants underpinning complex traits. Overall, however, this area of research looks set to deliver outcomes well into the future by capitalizing on new technological developments, complex multivariate analyses and the implementation of strategies to utilize new insights for breeding program improvement.

Author contribution statement BEH conceived of the idea and drafted the manuscript. KLV, VKS, PG, RKV, and CC prepared figures and contributed to the manuscript. APV, CR, and HL contributed to the manuscript. All authors revised the manuscript.

Acknowledgments Many thanks to three anonymous reviewers for their helpful suggestions. Dr. Huang is the recipient of an Australian Research Council Discovery Early Career Researcher Award (Project Number DE120101127).

Conflict of interest No authors have any conflicts of interest.

References

Ahfock D, Wood I, Stephen S, Cavanagh CR, Huang BE (2014) Characterizing uncertainty in high-density maps from multiparental populations. Genetics 198:117–128

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61

Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA et al (2011) Genetic analysis of complex traits in the emerging collaborative cross. Genome Res 21:1213–1222

Bailey DW (1971) Recombinant-inbred strains: an aid to finding identity, linkage, and function of histocompatibility and other genes. Transplantation 11:325–327

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT et al (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:11

Bardol N, Ventelon M, Mangin B, Jasson S, Loywick V et al (2013) Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (*Zea mays* L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. Theor Appl Genet 126:2717–2736

Bink MCAM, Boer MP, Ter Braak CJF, Jansen J, Voorrips RE et al (2008) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161:85–96

Blakeslee AF, Belling J, Farnham ME, Bergner AD (1922) A haploid mutant in the Jimson weed, "Datura Stramonium". Science 55:646–647

Bottomly D, Ferris MT, Aicher LD, Rosenzweig E, Whitmore A et al (2012) Expression quantitative trait loci for extreme host response to influenza A in pre-Collaborative Cross mice. G3 2:213–221

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Brim CA (1966) A modified pedigree method of selection in soybeans. Crop Sci 6:220

Broman K (2005) The genomes of recombinant inbred lines. Genetics 169:1133–1146

Broman KW (2012) Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. Genetics 190:403–412

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Buet C, Dubreuil P, Tixier M-H, Durantin K, Praud S et al (2013) The molecular characterization of a MAGIC population reveals high potential for gene discovery. MaizeGDB proceedings

Butler D (2009) asreml: asreml() fits the linear mixed model. R package version 3.0. http://www.vsni.co.uk

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Op Plant Biol 11:215–221

Cavanagh C, Chao S, Wang S, Huang BE, Stephen S et al (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci 110:8057–8062

Collaborative Cross Consortium (2012) The genome architecture of the collaborative cross mouse genetic reference population. Genetics 190:389–401

Complex Trait Consortium (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137

Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF (2013) Genetic incompatibilities are widespread within species. Nature 504:135–137

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11:381–393

Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141:1199–1207

Das S, Zijdenbos AP, Harlap J, Vins D, Evans AC (2011) LORIS: a web-based data management system for multi-center studies. Front Neuroinform 5:37

Demarest K, Koyner J, McCaughran J Jr, Cipp L, Hitzemann R (2001) Further characterization and high- resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. Behav Genet 31:79–91

Durrant C, Mott R (2010) Bayesian quantitative trait locus mapping using inferred haplotypes. Genetics 184:839–852

Durrant C, Swertz MA, Alberts R, Arends D, Moller S et al (2012) Bioinformatics tools and database resources for systems genetics analysis in mice—a short review and an evaluation of future needs. Brief Bioinform 13:135–142

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K et al (2011) A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379

Esch E, Szymaniak JM, Yates H, Pawlowski WP, Buckler ES (2007) Using crossover breakpoints in recombinant inbred lines to identify quantitative trait loci controlling the global recombination frequency. Genetics 177:1851–1858

Forster BP, Bors-Heberle E, Kasha KJ, Touraev A (2007) The resurgence of haploids in higher plants. Trends Plant Sci 12: 368–375

Furbank RT, Tester M (2011) Phenomics—technologies to relieve the phenotyping bottleneck. Trends Plant Sci 16:635–644

Gan X, Stegle O, Behr J, Steffen JG, Drewe P et al (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477:419–423

Gaur PM, Jukanti AK, Varshney RK (2012) Impact of genomic technologies on chickpea breeding strategies. Agronomy 2:199–221

Giraud H, Lehermeier C, Bauer E, Falque M, Segura V et al (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the Flint and Dent heterotic groups for maize. Genetics 198:1717–1734

Goulden CH (1939) Problems in plant selection. In: Proceedings of the Seventh International Genetics Congress. Cambridge University Press, pp 132–133

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

Gyenesei A, Moody J, Semple CAM, Haley CS, Wei W-H (2012) High throughput analysis of epistasis in genome-wide association studies with BiForce. Bioinformatics 28:1957–1964

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Harushima Y, Yano M, Shomura A, Sato M, Shimano T et al (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics 148:479–494

Hemani G, Theocharidis A, Wei W, Haley C (2011) EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 27:1462–1465

Hickey JM, Gorjanc G, Hearne S, Huang BE (2014) AlphaMPSim: flexible simulation of multi-parent crosses. Bioinformatics 30:2686–2688

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L et al (2008) Big data: the future of biocuration. Nature 455:47–50

Huang BE, George AW (2011) R/mpMap: a computational platform for the genetic analysis of multi-parent recombinant inbred lines. Bioinformatics 27:727–729

Huang X, Feng Q, Qian Q, Zhao Q, Wang L et al (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068–1076

Huang X, Paulo M-J, Boer M, Effgen S, Keizer P et al (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. PNAS 108(11):4488–4493

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ et al (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10:826–839

Huang BE, Clifford D, Cavanagh C (2013) Selecting subsets of genotyped experimental populations for phenotyping to maximize genetic diversity. Theor Appl Genet 126:379–388

Huang BE, Raghavan C, Mauleon R, Broman KW, Leung H (2014) Imputation of low-coverage genotyping-by-sequencing in multiparental crosses. Genetics 197:401–404

Jansen RC (1994) Controlling the type I and type II errors in mapping quantitative trait loci. Genetics 138:871–881

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216

Kass RE (1993) Bayes factors in practice. Statistician 42:551–560

Kass RE, Raftery AE (1995) Bayes factors. JASA 90:773–795

King EG, Macdonald SJ, Long AD (2012a) Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics 191:935–949

King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S et al (2012b) Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res 22:1558–1566

King EG, Sanderson BJ, McNeil CL, Long AD, Macdonald SJ (2014) Genetic dissection of the *Drosophila melanogaster* female head transcriptome reveals widespread allelic heterogeneity. PLoS Genet 10(5):e1004322

Klasen JR, Piepho H-P, Stich B (2012) QTL detection power of multi-parental RIL populations in Arabidopsis thaliana. Heredity 108:626–632

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM et al (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5(7):e1000551

Lai K, Lorenc MT, Edwards D (2012) Genomic databases for crop improvement. Agronomy 2:62–73

Leroux D, Rahmani A, Jasson S, Ventelon M, Louis F et al (2014) Clusthaplo: a plug-in for MCQTL to enhance QTL detection using ancestral alleles in multi-cross design. Theor Appl Genet 127:921–933

Mace ES, Hunt CH, Jordan DR (2013) Supermodels: sorghum and maize provide mutual insight into the genetics of flowering time. Theor Appl Genet 126:1377–1395

Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J et al (2014) An eight-parent Multiparent Advanced Generation Inter-Cross population for winter-sown wheat: creation, properties and validation. G3 4:1603–1610

Malosetti M, van Eeuwijk DA, Boer MP, Casas AM, Elia M et al (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

Malosetti M, Ribaut J-M, van Eeuwijk FA (2013) The statistical analysis of multi-environment data: modelling genotype-by-environment interaction and its genetic basis. Front Physiol 4:44

Maluszynski M, Kasha KJ, Szareiko I (2003) Published doubled haploid protocols in plant species. In: Doubled haploid production in crop plants, a manual. Kluwer Academic Publishers, Dordecht, pp 309–335

McClearn GE, Wilson JR, Meredith W (1970) The use of isogenic and heterogenic mouse stocks in behavioral research. In: Lindzey G, Thiessen D (eds) Contributions to behavior-genetic analysis: the mouse as a prototype. Appleton Century Crofts, New York, pp 3–22

McMullen MD, Kresovich S, Villeda HS, Bradbury PJ, Li H et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Meuwissen TH, Goddard ME (2001) Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33:605

Mohring J, Piepho H-P (2009) Comparison of weighting in two-stage analyses of series of experiments. Crop Sci 39:1977–1988

Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12:433–436

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A new method for fine-mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci USA 97:12649–12654

Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L et al (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 (in press)

Pea G, Dell'Acqua M, Hlaing ALL, Pe ME (2013) From mice to maize: a multiparental population for fine mapping in *Zea mays*. MAGIC Populations Workshop. http://openwetware.org/images/e/e6/MatteoDellAcqua_MaizePoster.pdf

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Ram R, Mehta M, Balmer L, Gatti DM, Morahan G (2014) Rapid identification of major effect genes using the Collaborative Cross. Genetics 198:75–86

Rebetzke GJ, Verbyla AP, Verbyla KL, Morell MK, Cavanagh CR (2014) Use of a large multiparent wheat mapping population in genomic dissection of coleoptile and seedling growth. Plant Biotechnol J 12:219–230

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 Breeding 35:86

Schmitt CP, Burchinal M (2011) Data management practices for collaborative research. Front Psychiatry 2:47

Schnaithmann F, Kopahnke D, Pillen K (2014) A first step toward the development of a barley NAM population and its utilization to detect QTLs conferring leaf rust seedling resistance. Theor Appl Genet 127:1513–1525

Scutari M, Howell P, Balding DJ, Mackay IJ (2014) Multiple quantitative trait analysis using Bayesian networks. Genetics 198:129–137

Smith AB, Lim P, Cullis BR (2006) The design and analysis of multi-phase plant breeding experiments. J Agric Sci Camb 144:393–409

Smith AB, Thompson R, Butler DC, Cullis BR (2011) The design and analysis of variety trials using mixtures of composite and individual plot samples. J Royal Stat Soc C 60:437–455

Smith AB, Butler DG, Cavanagh CR, Cullis BR (2015) Multi-phase variety trials using both composite and individual replicate samples: a model-based design approach. J Agric Sci Camb (in press)

Stich B (2009) Comparison of mating designs for establishing Nested Association Mapping populations in maize and *Arabidopsis thaliana*. Genetics 183:1525–1534

Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R et al (2012) High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics 190:437–447

Thépot S, Restoux G, Goldringer I, Hospital F, Gouache D, Mackay I, Enjalbert J (2015) Efficiently tracking selection in a multiparental population: the case of earliness in wheat. Genetics 199:609–623

Valdar W, Flint J, Mott R (2006) Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics 172:1783–1797

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

van Eeuwijk FA, Bink MC, Chenu K, Chapman SC (2010) Detection and use of QTL for complex traits in multiple environments. Curr Opin Plant Biol 13:193–205

Verbyla AP, Cullis BR (2012) Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. Theor Appl Genet 125:933–953

Verbyla AP, Cullis BR, Thompson R (2007) The analysis of QTL by simultaneous use of the full linkage map. Theor Appl Genet 116:95–111

Verbyla AP, George AW, Cavanagh CR, Verbyla KL (2014a) Whole genome QTL analysis for MAGIC. Theor Appl Genet 127:1753–1770

Verbyla AP, Cavanagh CR, Verbyla KL (2014b) Whole genome analysis of multi-environment or multi-trait QTL in MAGIC G3(4):1569–1584

Wang J, de Villena FP, Lawson HA, Cheverud JM, Churchill GA et al (2012) Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny. Genetics 190:449–458

Wang S, Wong D, Forrest K, Allen A, Chao S et al (2014) Characterization of polyploidy wheat genomic diversity using the high-density 90,000 SNP array. Plant Biotechnol J 12:787–796

Xu S (1996) Mapping quantitative trait loci using four-way crosses. Genet Res 68:175–181

Yalcin B, Flint J, Mott R (2005) Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. Genetics 171:673–681

Yamamoto E, Iwata H, Tanabata T, Mizobuchi R, Yonemaru J et al (2014) Effect of advanced intercrossing on genome structure and on the power to detect linked quantitative trait loci in a multi-parent population: a simulation study in rice. BMC Genet 15:50

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

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468