Brachi *et al. Genome Biology* 2011, **12**:232
http://genomebiology.com/2011/12/10/232

Genome Biology

REVIEW

Genome-wide association studies in plants: the missing heritability is in the field

Benjamin Brachi, Geoffrey P Morris and Justin O Borevitz*

Abstract

Genome-wide association studies (GWAS) have been even more successful in plants than in humans. Mapping approaches can be extended to dissect adaptive genetic variation from structured background variation in an ecological context.

Introduction

The genetic sources of phenotypic variation have been a major focus of both plant and animal studies aimed at identifying the causes of disease, improving agriculture and understanding adaptive processes. In plants, quantitative trait loci (QTL) were originally mapped in bi-parental crosses, but they were restricted in allelic diversity and in having limited genomic resolution [1]. The genome-wide association approach (GWAS) overcomes several limitations of traditional gene mapping by (i) providing higher resolution, often to the gene level, and (ii) using samples from previously well-studied populations in which commonly occurring genetic variations can be associated with phenotypic variation. The advent of high-density single-nucleotide polymorphism (SNP) typing allowed whole-genome scans to identify often small haplotype blocks that are significantly correlated with quantitative trait variation. These approaches have enabled both large studies of human disease, which have identified important loci [2], and recent plant studies that have been successful in identifying loci that explain large portions of phenotypic variation.

Significant associations between genetic variations and phenotypic diversity have been found in some human studies, but they explain only a few percent of the phenotypic diversity, leading many geneticists to ask 'Where is the missing heritability?' [3,4]. This question has several possible answers. First, rare variants [3-5], major alleles that are unique to local families, can be detected only when sampling is adequate at the local level. Second, allelic heterogeneity, the phenomenon in which multiple functional alleles of the same gene exist and are associated with different phenotypes, is common, especially in wide population samples [6-8]. Third, single-marker approaches suffer from genetic heterogeneity when multiple major loci are involved and in linkage disequilibrium (LD) with each other [9]. Fourth, variation resulting from epistatic interactions between genes might go undiscovered because epistasis can only be investigated practically in a sequential scan of major common loci and the genome [10]. Finally, epigenetic variation, which requires sophisticated genotyping, is likely to be a source of missing heritability [11]. The influence of each of these factors on heritability strongly depends on the population sampled. Thus, even true positives will often fail to replicate across populations. Owing to the confounding effect of population structure, true causative SNPs are difficult to identify because they are in LD (that is, in non-random association) with many loci in the genome [6].

When human GWAS find associations that have genome-wide significance, the SNPs explain only a tiny fraction of the phenotypic variation revealed by family-based studies [12]. But the results of recent GWAS in plants (in *Arabidopsis thaliana*, rice, and maize) have explained a much greater proportion of the phenotypic variation than that explained by human GWAS studies. It seems that, in plants at least, the assumption that common genetic variation explains common phenotypic variation holds. In plants, rare variation can become sufficiently common in large families or populations to be identifiable by GWAS. For example, GWAS have identified SNPs and population structure that can explain up to 45% of the phenotypic variation in flowering time [13]. However, flowering time has even higher heritability (approximately 90%), leaving an additional 45% of heritable variation unexplained.

In this review, we consider why GWAS in plants have been successful, focusing on the experimental designs and sampling strategies used in these studies. Those working on GWAS in human genetics and in plants have

*Correspondence: borevitz@uchicago.edu
Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

much to learn from each other. We then discuss future developments for generalized GWAS in plants, taking on board the lessons learned in model species. Empirical geographic knowledge of gene flow and population structure, together with hypotheses about the ecological zones that have imposed selection, enables the sampling of different populations in which the same or different adaptive traits are inherited. A general population restructuring approach can then be used to uncouple adaptive variation from the genomic background through synthetic outcrossing among lines that have balanced genetic diversity.

GWAS in model species

Arabidopsis thaliana

Finding the genetic basis of complex traits in plants, such as flowering time, growth rate, and yield, has been a major focus of attempts to improve crops and understand plant adaptation. *A. thaliana* has long been an attractive model for the study of natural variation and adaptation because of its wide distribution [14], the diversity of its habitats, and the unequaled genomic resources available for this species. GWAS requires a genomic map in which the marker density is higher than the extent of LD. This, in turn, depends on the population sample, specifically the standing genetic diversity and the number of recombination events that shuffle that diversity. In a global set of *A. thaliana* accessions, LD was shown to decay within 10 kb on average, so the optimal number of SNPs necessary to cover the whole genome was estimated to be 140,000 [15]. A genotyping array, designed to type 250,000 SNPs, was used to genotype an initial set of 192 natural accessions [16]. In this seminal study, an extensive set of 107 phenotypes were used to run GWAS in *A. thaliana*. To test the ability of GWAS to detect the genetic basis of natural variation efficiently, the power to detect previously identified candidate genes was assessed through the calculation of enrichment ratios. In most cases, large enrichment ratios were found [16,17], meaning that SNPs with high association scores were more likely to be close to previously identified candidate genes than random loci. Furthermore, some of the alleles identified in GWAS overlapped with lower-resolution QTL identified with recombinant inbred line (RIL) mapping [13,17,18]. Together, this evidence conclusively demonstrates that GWAS can identify many true genotype-phenotype associations.

The potential of GWAS in *A. thaliana* was demonstrated by the successful functional validation of the gene *ACCELERATED CELL DEATH6* (*ACD6*) [19]. Natural variation in *ACD6* was shown to underpin differences in vegetative growth and in resistance to microbial infection and herbivory [20]. A Col-0 (reference accession) background with a loss-of-function allele of *ACD6* displayed increased leaf necrosis, reduced growth and reduced susceptibility to different pathogens when transformed with the *ACD6* allele from the Est-1 accession. GWAS was performed for leaf necrosis on a set of 96 natural accessions. Nine of the fifteen SNPs with the lowest *P*-values in this scan were located close to or within *ACD6*. None of the new genes identified by GWAS have been functionally validated to date, but this study confirms the ability of GWAS to detect true positives as *ACD6* was previously known from forward-genetic mutant screens [20].

Allowing for the average LD distance (10 kb) is sufficient to enable the identification of individual genes, but the gene density seen in *A. thaliana* suggests that some genomic regions display extended clusters of high-scoring SNPs instead of sharp peaks. The broad 'mountain range' of associations makes the selection of candidate genes difficult [16]. The width of the 'mountain' can be broad due to extended LD from a recent selective sweep or because of low recombination. In addition, genetic or allelic heterogeneity can create 'mountain ranges' that have multiple peaks. The sweeps acting on common loss-of-function deletions at *FRIGIDA* (*FRI*), along with other linked flowering time loci, probably explain the complex pattern of association with flowering time that was observed at this locus [16]. Tightly linked genes have been shown to underlie a complex association with growth rate variation [21]. Another limitation to the ability of GWAS to identify individual genes is the occurrence of false positives that are an artifact of population structure [22]. The worldwide set of natural *A. thaliana* accessions is highly structured [23], and when phenotypic variation for the trait of interest overlaps with patterns of population structure, strong confounding can occur. Statistical methods that have been developed to control for population structure [21,24-27] produce a *P*-value distribution that is closer to a uniform distribution, although they can have reduced sensitivity. Nevertheless, GWAS in *A. thaliana* have been shown to have significant power in detecting previously known candidate genes, and they have also detected hundreds of loci that are involved in the natural variation of complex traits. This new knowledge of the number of genes that underlie adaptive traits, and the size of their effects, allows us to better understand the bases of flowering time, growth rate, and yield.

Maize and rice

Maize and rice, two of the most important crop species in the world, have been the focus of intense efforts to map the ancestral genetic variation that underlies agronomic traits such as grain yield, disease resistance, and plant architecture. Maize is an outcrossing plant, with an LD that decays at approximately 2,000 bp on average (a

distance 5-fold shorter than that in *A. thaliana* [28]). It also has a large genome (2.3 Gb of unique sequence [29]), and thus the typing of many SNPs is required to define a haplotype map for maize. A set of 1.6 million SNPs has been designed for maize GWAS, but the dense genotyping of a large number of lines was initially prohibited by cost.

The approach that was taken instead was to genotype a limited number of lines (25 founders) and to cross them to produce 25 RIL families, known as the nested association mapping (NAM) populations [30]. A total of 5,000 RILs (200 per family) were then genotyped at low density. High-density genotypes were then imputed on the basis of high-density genotypes of the founding lines. The complete set of RILs was phenotyped, and SNP associations were then tested across all the RILs, with the test including a term to account for variation caused by the RIL family effect. The main advantages of this approach are: (i) the imputation of high-density genotypes gives some fine-mapping resolution among the 25 founders; (ii) outcrossing reshuffles variation in the founder genomes and therefore provides some control of population structure effects; (iii) joint-linkage mapping identifies low-resolution QTL across all RIL families, and this genetic background can be controlled while performing nested associations for fine mapping; and (iv) the use of RILs allows repeated measures of phenotypes on the same lines, in common and different environments, allowing precise estimation of variation in traits such as flowering time [31], leaf architecture [32], and blight resistance [33,34]. NAM also has some limitations, primarily that the small number of founders limits genetic diversity and ancestral recombination. One special strength is that high-density genotypes are imputed onto progeny typed with fewer markers, where new recombinations have shuffled SNPs that were previously in LD because of population structure. Many designs of NAM are likely to emerge that fit the particular population biology of the target species [35].

Rice is a selfing species and, like *A. thaliana*, a good candidate for GWAS. Huang *et al.* [36] identified an unbiased set of common SNPs that they used to identify strong associations between genetic loci and 14 agronomic traits, including heading date, grain size, and starch quality. Here, the step forward was to use a strategy based on second-generation sequencing technology to develop a haplotype map for 517 Chinese land races across the *Oryza indica* and *Oryza japonica* rice subspecies. The idea was to perform low depth (1X) whole-genome sequencing, and then take advantage of the >100 kb LD in rice to impute missing data. This strategy was successful because the imputation algorithm that was developed reduced the missing data from 60% to 3%, with 98% accuracy. GWAS was subsequently performed using 671,355 SNPs in a subset of 373 *indica* lines to avoid the major confounding of population structure between subspecies. This identified between 1 and 7 loci for each agronomic trait, each of which explained between 6% and 68% of the variation in that trait. A few genes that have large effects in controlling traits that are involved in determining yield, morphology, stress tolerance, and nutritional quality were also identified in recent rice GWAS [37,38]. Together, these studies establish a research platform that can link genomic variation and germplasm collections to enable molecular breeding.

The scale of adaptive variation

Controlling for population structure is a standard procedure in GWAS, although doing this when the traits are strongly confounded reduces the power of the analysis and can lead to false negatives. This issue is especially likely to arise when studying traits such as flowering time and cold tolerance, which are filtered by environmental gradients that overlap with patterns of population structure. In this case, controlling for population structure can reduce the association signals around major adaptive genes [6,17,39]. In this situation, the only solution is synthetic, that is, to re-structure populations by making crosses. Another weakness of GWAS is its lack of power to detect rare alleles that are involved in natural variation. Parametric tests of association, including efficient mixed-model association (EMMA) [40,41], are sensitive to SNPs that have low minor-allele frequencies, which can show an artificially increased association score (-log(*P*-value)). Because of this phenomenon, most studies have not considered SNPs that have minor allele frequencies under 5% or 10%, although these variants do contribute to phenotypic variation [24]. Balancing samples across population subdivisions can homogenize allele frequencies, elevating globally rare variants that are common in certain subdivisions. Their direct trait association can be detected when they are decoupled from population structure. Allelic heterogeneity is another limitation that applies to GWAS and other multi-parent mapping strategies [42,43] because GWAS assumes that common (biallelelic) genetic variation explains quantitative trait variation [6,17]. Association tests involving SNPs that tag multiple alleles in LD with each other can therefore be positively misleading [9].

Some of the confounding effects of population structure in GWAS can be avoided by adjusting the sampling strategy (Figure 1). Characterization of population structure before carrying out the GWAS, along with knowledge of the ecological factors that are imposing selection, will help to address certain pitfalls of GWAS and will enable the dissection of adaptive variation from structured background variation. A theoretical example

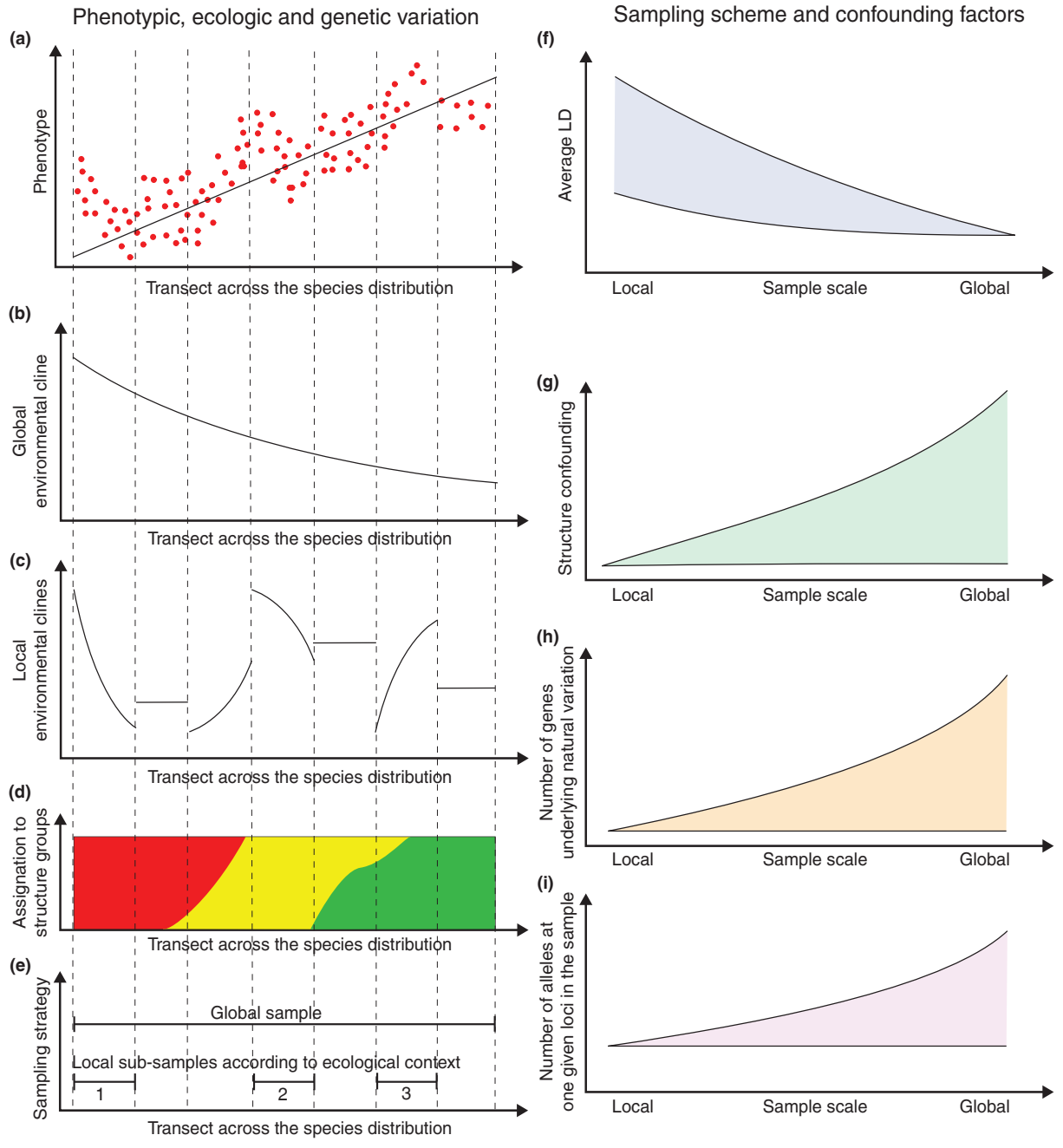

Figure 1. Influence of sampling strategy on GWAS confounding effects. (a) Relationship between an adaptive trait and the position along a transect across the species distribution. The phenotype could, for example, be flowering time in *A. thaliana*, and accession lines could have been sampled along a transect from the south to the north of the species' distribution. The relationship is positive because the phenotype is adaptive to an environmental variable varying along this transect. **(b)** Some traits show a gradual change along the transect. In the example of flowering time in *A. thaliana*, environmental factors such as temperature and photoperiod would show continuous change along the latitudinal clines. But **(c)** the phenotypes also show extensive variation at a given position along the transect, suggesting that other ecological factors, acting at smaller scales, might also be acting as selective pressures on the phenotype. These local environmental variations could be related to soil quality, exposition, competition or predation. They can differ between sites that are close to one another without following a trend across the entire species' distribution range. **(d)** The genetic structure of a species can be represented as the proportion of individuals assigned to each of three structure groups along the species-wide transect. **(e)** A global sample covers the entire species repartition range; alternatively, local sub-samples can be taken at locations chosen with reference to the pattern of the population structure and to small-scale environmental variations that have the potential to act as selective pressures. **(f-i)** Effect of the sampling scale (from local to species-wide sampling) on LD and confounding factors.

is presented in Figure 1, where we also discuss the sampling of *A. thaliana* accessions and the confounding factors that population sampling can bring to GWAS. Patterns of population structure overlap with patterns of the phenotype and with patterns of environmental variation (Figure 1a,b,d), increasing the rate of false positives and false negatives in GWAS. Variation in flowering time across widely distributed accessions may involve many genes and even multiple alleles of those genes. If adaptive alleles are locally common but globally rare, a broad sample will have low power to find significant phenotypic associations. At a smaller spatial scale (for example, the local population level), the phenotypic variation is largely restricted to the variation present in the founders. At this scale, there is increased statistical power to detect the more limited genetic diversity (Figure 1h). By contrast, in a species-wide sample, the loci that underlie natural variation might harbor multiple rare alleles (Figure 1i) that are likely to go undetected, and hence the power of association scans is decreased. Local environmental variations are also likely to apply selective pressures on adaptive traits, explaining some of the variation around the global trend seen at the species scale (Figure 1c). In the case of flowering time in *A. thaliana,* these ecological factors could be soil composition, slope, aspect, sun exposure, or even biotic factors such as herbivory and pathogens. Choosing multiple sub-samples (Figure 1d,e) to avoid major confounding by population structure (Figure 1g) is advantageous, but using sub-samples from locations where phenotypic variation spans ecological conditions makes it possible to map adaptive variation within a largely unstructured set (Figure 1c,e).

The current collection of more than 1,300 *A. thaliana* accessions, genotyped at 250,000 SNPs (M Horton, J Bergelson and M Nordborg, personal communication) and eventually the data from the 1001 Genomes Project [44], are large enough samples to begin to deliver empirical knowledge of the deeper patterns of genomic variation on the landscape [45]. By gleaning the genetic information, one can select a core mapping subset, like the RegMap lines in *A. thaliana* [6], that has balanced regional diversity and reduced confounding effects of population structure, but an average length of LD decay that is short enough to allow precise mapping of the underlying genes. The distribution of some phenotypes might, however, overlap with patterns of population structure at a local scale. For example, this could be the case in a newly colonized region where a patchy distribution offers little opportunity for gene flow. Large parts of the genome (or the whole genome with complete isolation) might be selected along with the genes controlling locally adaptive phenotypes. In this case, approaches involving wider crosses seem to be the only way to identify the underlying genes.

The scale of genomic variation

New genotyping-by-sequencing (GBS) technologies and bioinformatic methods, based on light shotgun sequencing or reduced representation and multiplexing, have the ability to discover, genotype, and impute near-complete population genomic data in any species [46-49]. For a given sequencing investment, there is a trade-off between the sample sequencing depth and the number of samples; with multiplexing, more samples can be sequenced but with lighter coverage. Importantly for imputation, the sequencing depth required for each individual depends on the extent of LD. The increased LD within families allows the haplotype map to be imputed from lower-coverage data, and this is an important advantage of the NAM design [30]. With moderate LD, the rice haplotype map could be assembled from hundreds of landraces typed at 1X coverage [36]. To integrate linkage-based pedigrees and association studies, GBS can be used to type progeny from several maternal lines of population samples. As has been achieved in rice, high-resolution genotypes of the maternal line could be assembled and near-complete genotypes imputed for the progeny.

Genotyping arrays only include a fraction of the SNPs identified in a restricted set of lines. Some missing heritability probably originates from the characterization of the genetic diversity using ascertained SNPs, which reduces the ability to detect rare alleles and causal polymorphisms. This can lead to an underestimation of the diversity and relatedness [50]. This component of missing heritability can be largely overcome by next generation sequencing technology, but repetitive and highly divergent portions of the genome might remain largely inaccessible. Aligning short reads to a single reference genome might introduce some ascertainment bias but this should be less of an issue as reads become longer.

An emerging synthesis for adaptation genetics: finding the missing heritability

The study of adaptation in traditional model plants such as *A. thaliana*, maize, and rice has been moving back 'into the field', with new wild collections and greater ecological context being introduced. At the same time, model systems of plant adaptation, such as columbine (*Aquilegia*) [51], monkey flower (*Mimulus*) [52], and sunflower (*Helianthus*) [53], can now take advantage of genomic tools that enable association mapping. This convergence of disciplines points towards an emerging synthesis of adaptation genetics.

We can suggest strategies to look for the missing heritability using genome-wide association mapping, but the optimal strategy will depend on the trait of interest and the scale at which it is adaptive. The first step will often be to use a large, geographically wide, and hierarchically structured starting sample to characterize

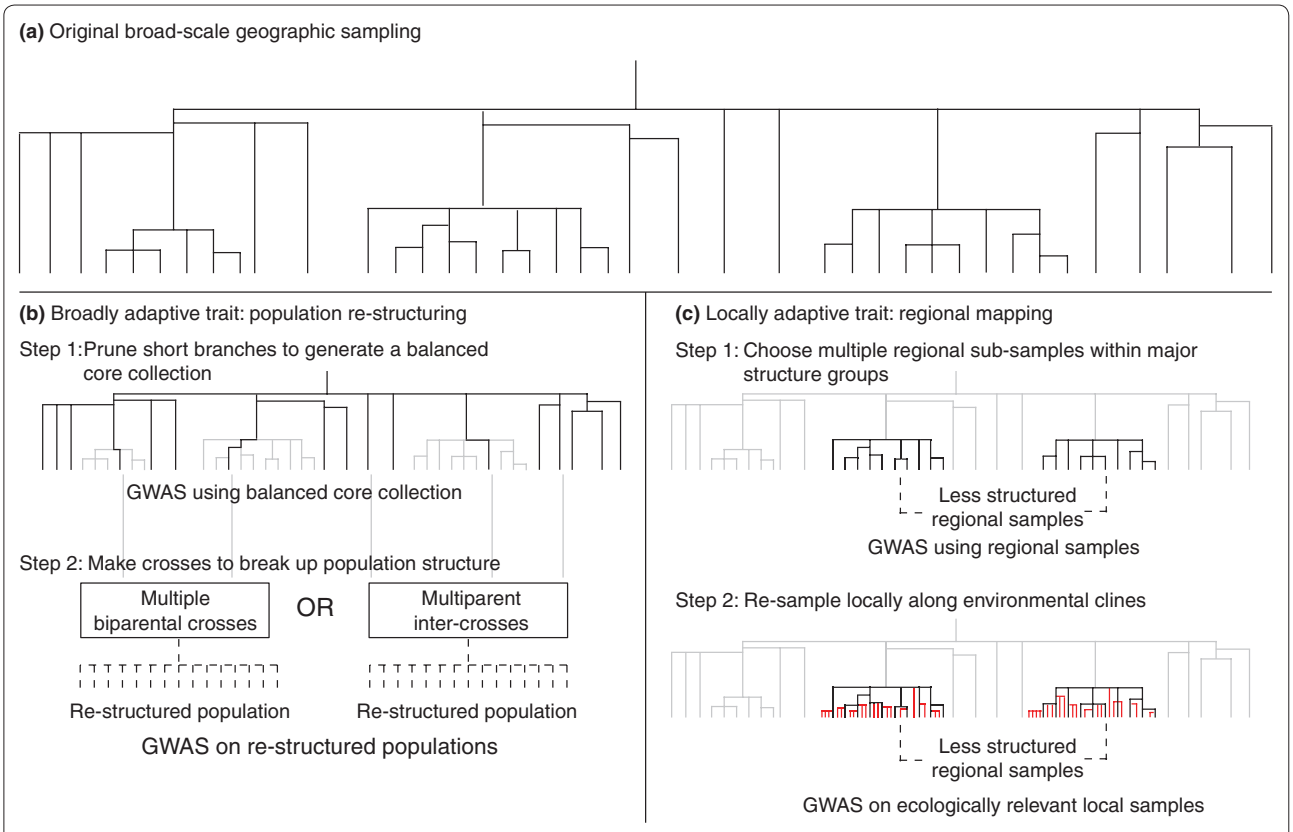

Figure 2. Strategies for GWAS include population re-structuring and regional re-sampling. (a) A schematic phylogenetic tree illustrating genetic diversity and population structure in a hypothetical sample of a species whose adaptive traits are to be investigated genetically. **(b)** To map the loci underlying adaptation at the broadest scale, a balanced core set of accessions is made by pruning closely related individuals from the global set. GWAS can be performed at this stage, but for traits whose variation that is confounded by population structure (Figure 1), crosses are needed. **(c)** To map loci underlying local adaption, the focus should be on less structured regional sub-samples that are identified in the initial sample (for example, RegMap lines in *A. thaliana*). GWAS can be performed on these regional samples, which have reduced allelic heterogeneity and confounding by population structure, but LD blocks are likely to be longer in the regional subsets and this will decrease mapping resolution. Regional re-sampling along an environmental cline in the field can increase the power of the association mapping.

population structure and gene flow empirically using non-ascertained markers (Figure 2a). If the trait of interest is adaptive on a broad scale, as is often assumed for traits that display latitudinal variation [54-58], then the genes should be mapped using accessions from across the range. At this scale, however, confounding is likely to occur when patterns of phenotypic variation overlap with patterns of population structure. In this case, 'population re-structuring' should be applied. This uses multiple crosses between a balanced set of diverse founders to break up population structure, while taking advantage of short LD blocks from ancestral recombination. The underlying principal is similar to combined GWAS and admixture mapping of the human genome, which takes advantage of both ancient and recent recombination events [59]. Several population re-structuring approaches have been used to date [30,36,43]; all require high-throughput phenotyping strategies to phenotype a very large number of lines. Regional sub-sampling within population structure groups can be performed to identify loci that are involved in adaptation to local environmental variations (Figure 2c). To improve power and resolution, re-sampling can be performed along the target environmental cline to increase the sample size within structure groups (Figures 1 and 2c).

Studies in model species such as *A. thaliana*, rice and maize have validated these approaches to identifying the genetic bases of adaptive traits. These methods, when combined with the increasing capacity and decreasing costs of next-generation sequencing, will allow GWAS in non-model species. Ultimately, population genomic studies across multiple species that occupy the same habitats will allow comparative studies of adaptive genetic variation among species that have potentially evolved in parallel under the same selective pressures. A better understanding of adaptive processes at the community level might be obtained by comparing the genetic architectures of adaptive traits among species that may

have different life histories. We believe that developing landscape and population genomic resources together in new species will enable high-power association mapping experiments to find the missing heritability underlying the adaptive traits seen in the field.

Acknowledgements
We thank Paul Grabowski, Alex Platt, and Magnus Nordborg for discussion about the manuscript and the reviewers for their comments. BB is supported by a Dropkin fellowship and an NIH grant to Joy Bergelson; GPM is supported by the Argonne-University of Chicago Strategic Collaborative Initiative, and JOB is supported by NIH RO1 GM073822.

Published: 28 October 2011

References

1. Borevitz JO, Nordborg M: **The impact of genomics on the study of natural variation in *Arabidopsis*.** *Plant Physiol* 2003, **132:**718-725.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.
3. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease.** *Nat Rev Genet* 2010, **11:**446-450.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.
5. Luo L, Boerwinkle E, Xiong M: **Association studies for next-generation sequencing.** *Genome Res* 2011, **21:**1099-1108.
6. Bergelson J, Roux F: **Identifying the genetic basis of complex traits in *Arabidopsis thaliana*.** *Nat Rev Genet* 2010, **11:**867-879.
7. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, Harries LW, Chong S, Moore M, Weedon MN, Guralnik JM, Bandinelli S, Murray A, Ferrucci L, Singleton AB, Melzer D, Frayling TM: **Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association.** *Hum Mol Genet* 2011, **20:**4082-4092.
8. Zhang X, Cal AJ, Borevitz JO: **Genetic architecture of regulatory variation in *Arabidopsis thaliana*.** *Genome Res* 2011, **21:**725-733.
9. Platt A, Vilhjálmsson BJ, Nordborg M: **Conditions under which genome-wide association studies will be positively misleading.** *Genetics* 2010, **186:**1045-1052.
10. Storey JD, Akey JM, Kruglyak L: **Multiple locus linkage analysis of genome-wide expression in yeast.** *PLoS Biol* 2005, **3:**e267.
11. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P, Bouchez D, Dillmann C, Guerche P, Hospital F, Colot V: **Assessing the impact of transgenerational epigenetic variation on complex traits.** *PLoS Genetics* 2009, **5:**e10000530.
12. Ott J, Kamatani Y, Lathrop M: **Family-based designs for genome-wide association studies.** *Nat Rev Genet* 2011, **12:**465-474.
13. Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO: **Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*.** *Proc Natl Acad Sci U S A* 2010, **107:**21199-21204.
14. Shindo C, Bernasconi G, Hardtke CS: **Natural genetic variation in *Arabidopsis*: tools, traits and prospects for evolutionary ecology.** *Ann Bot* 2007, **99:**1043-1054.
15. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M: **Recombination and linkage disequilibrium in *Arabidopsis thaliana*.** *Nat Genet* 2007, **39:**1151-1155.
16. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JD, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, *et al.*: **Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines.** *Nature* 2010, **465:**627-631.
17. Brachi B, Nathalie F, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F: **Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature.** *PLoS Genet* 2010, **6:**e1000940.
18. Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JDG: **Genome-wide survey of *Arabidopsis* natural variation in downy mildew resistance using combined association and linkage mapping.** *Proc Natl Acad Sci U S A* 2010, **107:**10302-10307.
19. Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, Laitinen RA, Huang Y, Chory J, Lipka V, Borevitz JO, Dangl JL, Bergelson J, Nordborg M, Weigel D: **Natural allelic variation underlying a major fitness trade-off in *Arabidopsis thaliana*.** *Nature* 2010, **465:**632-636.
20. Lu H, Rate DN, Song JT, Greenberg JT: **ACD6, a novel ankyrin protein, is a regulator and an effector of salicylic acid signaling in the *Arabidopsis* defense response.** *Plant Cell* 2003, **15:**2408-2420.
21. Kroymann J, Mitchell-Olds T: **Epistasis and balanced polymorphism influencing complex trait variation.** *Nature* 2005, **435:**95-98.
22. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:**945-959.
23. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J: **The pattern of polymorphism in *Arabidopsis thaliana*.** *PLoS Biol* 2005, **3:**e196.
24. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costish DE, Buckler ES: **Association mapping: critical considerations shift from genotyping to experimental design.** *Plant Cell* 2009, **21:**2194-2202.
25. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES: ***Dwarf8* polymorphisms associated with variation in flowering time.** *Nature Genet* 2001, **28:**286-289.
26. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nat Genet* 2006, **38:**203-208.
27. Zhao K, Nordborg M, Marjoram P: **Genome-wide association mapping using mixed-models: application to *GAW15* Problem 3.** *BMC Proc* 2007, **1 Suppl 1:**S164.
28. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in maize genome.** *Proc Natl Acad Sci U S A* 2001, **98:**11479-11484.
29. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al.*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112-1115.
30. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, *et al.*: **Genetic properties of the maize nested association mapping population.** *Science* 2009, **325:**737-740.
31. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N, *et al.*: **The genetic architecture of maize flowering time.** *Science* 2009, **325:**714-718.
32. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES: **Genome-wide association study of leaf architecture in the maize nested association mapping population.** *Nat Genet* 2011, **43:**159-162.
33. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB: **Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population.** *Nat Genet* 2011, **43:**163-168.
34. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ: **Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize.** *Proc Natl Acad Sci U S A* 2010, **108:**6893-6898.

35. Stich B: **Comparison of mating designs for establishing nested association mapping populations in maize and *Arabidopsis thaliana*.** *Genetics* 2009, **183:**1525-1534.
36. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B: **Genome-wide association studies of 14 agronomic traits in rice landraces.** *Nat Genet* 2010, **42:**961-967.
37. Famoso AN, Zhao K, Clark RT, Tung C-W, Wright MH, Bustamante C, Kochian LV, McCouch SR: **Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping.** *PLoS Genet* 2011, **7:**e1002221.
38. Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR: **Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*.** *Nat Commun* 2011, **2:**467.
39. Mitchell-Olds T, Schmitt J: **Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*.** *Nature* 2006, **441:**947-952.
40. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet* 2010, **42:**348-354.
41. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178:**1709-1723.
42. Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA: **Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population.** *Proc Natl Acad Sci* 2011, **108:**4488-4493.
43. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R: **A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*.** *PLoS Genet* 2009, **5:**e1000551.
44. **1001 Genomes: a Catalog of *Arabidopsis thaliana* Genetic Variation** [http://1001genomes.org/]
45. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Agren J, Bossdorf O, Byers D, Donohue K, Dunning M, Holub EB, Hudson A, Le Corre V, Loudet O, Roux F, Warthmann N, Weigel D, Rivero L, Scholl R, Nordborg M, Bergelson J, Borevitz JO: **The scale of population structure in *Arabidopsis thaliana*.** *PLoS Genet* 2010, **6:**e10000843.
46. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL: **Multiplexed shotgun genotyping for rapid and efficient genetic mapping.** *Genome Res* 2011, **21:**610-617.
47. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.** *PLoS ONE* 2011, **6:**e19379.
48. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA: **Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.** *PLoS Genet* 2010, **6:**e1000862.
49. Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q: **Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing.** *Proc Natl Acad Sci U S A* 2010, **107:**10578-10583.
50. Rosenblum EB, Novembre J: **Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard.** *J Heredity* 2007, **98:**331-336.
51. Kramer EM, Hodges SA: ***Aquilegia* as a model system for the evolution and ecology of petals.** *Philos Trans R Soc Lond B Biol Sci* 2010, **365:**477-490.
52. Wu CA, Lowry DB, Cooley AM, Wright KM, Lee YW, Willis JH: ***Mimulus* is an emerging model system for the integration of ecological and genomic studies.** *Heredity* 2007, **100:**220-230.
53. Strasburg JL, Kane NC, Raduski AR, Bonin Al, Michelmore R, Rieseberg LH: **Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers.** *Mol Biol Evol* 2011, **28:**1569-1580.
54. Ducrocq S, Veyrieras J-B, Camus-Kulandaivelu L, Kloiber-Maitz M, Presterl T, Ouzunova M, Manicacci D, Charcosset A: **Key impact of *Vgt1* on flowering time adaptation in maize: evidence from association mapping and ecogeographical information.** *Genetics* 2008, **178:**2433-2437.
55. Jones H, Leigh FJ, Mackay I, Bower MA, Smith LMJ, Charles MP, Jones G, Jones MK, Brown TA, Powell W: **Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent.** *Mol Biol Evol* 2008, **25:**2211-2219.
56. Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J: **A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*.** *Proc Natl Acad Sci U S A* 2004, **101:**4712-4717.
57. Uga Y, Nonoue Y, Liang Z, Lin H, Yamamoto S, Yamanouchi U, Yano M: **Accumulation of additive effects generates a strong photoperiod sensitivity in the extremely late-heading rice cultivar 'Nona Bokra'.** *Theor Appl Genet* 2007, **114:**1457-1466.
58. Van Dijk H, Boudry P, McCombie H, Vernet P: **Flowering time in wild beet (*Beta vulgaris* ssp. *maritima*) along a latitudinal cline.** *Acta Oecologica* 1997, **18:**47-60.
59. Seldin MF, Pasaniuc B, Price AL: **New approaches to disease mapping in admixed populations.** *Nat Rev Genet* 2011, **12:**523-528.

doi:10.1186/gb-2011-12-10-232
Cite this article as: Brachi B, *et al*.: Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* 2011, **12:**232.