LETTERS

Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines

Susanna Atwell[1]*, Yu S. Huang[1]*, Bjarni J. Vilhjálmsson[1]*, Glenda Willems[1]*, Matthew Horton[3], Yan Li[3], Dazhe Meng[1], Alexander Platt[1], Aaron M. Tarone[1], Tina T. Hu[1], Rong Jiang[1], N. Wayan Muliyati[3], Xu Zhang[3], Muhammad Ali Amer[1], Ivan Baxter[4], Benjamin Brachi[6], Joanne Chory[7,8], Caroline Dean[9], Marilyne Debieu[10], Juliette de Meaux[10], Joseph R. Ecker[8], Nathalie Faure[6], Joel M. Kniskern[3], Jonathan D. G. Jones[11], Todd Michael[8], Adnane Nemri[11], Fabrice Roux[3,6], David E. Salt[5], Chunlao Tang[1], Marco Todesco[12], M. Brian Traw[3], Detlef Weigel[12], Paul Marjoram[2], Justin O. Borevitz[3], Joy Bergelson[3] & Magnus Nordborg[1,13]

Although pioneered by human geneticists as a potential solution to the challenging problem of finding the genetic basis of common human diseases[1,2], genome-wide association (GWA) studies have, owing to advances in genotyping and sequencing technology, become an obvious general approach for studying the genetics of natural variation and traits of agricultural importance. They are particularly useful when inbred lines are available, because once these lines have been genotyped they can be phenotyped multiple times, making it possible (as well as extremely cost effective) to study many different traits in many different environments, while replicating the phenotypic measurements to reduce environmental noise. Here we demonstrate the power of this approach by carrying out a GWA study of 107 phenotypes in *Arabidopsis thaliana*, a widely distributed, predominantly self-fertilizing model plant known to harbour considerable genetic variation for many adaptively important traits[3]. Our results are dramatically different from those of human GWA studies, in that we identify many common alleles of major effect, but they are also, in many cases, harder to interpret because confounding by complex genetics and population structure make it difficult to distinguish true associations from false. However, a-priori candidates are significantly over-represented among these associations as well, making many of them excellent candidates for follow-up experiments. Our study demonstrates the feasibility of GWA studies in *A. thaliana* and suggests that the approach will be appropriate for many other organisms.

The genotyped sample (Supplementary Table 1) includes a core set of 95 lines[4] for which a wide variety of phenotypes were available, plus a second set of 96 lines for which many phenotypes related to flowering were available[5]. The lines were genotyped using a custom Affymetrix single nucleotide polymorphism (SNP) chip containing 250,000 SNPs[6]. Because the genome of *A. thaliana* is around 120 megabases long and the extent of linkage disequilibrium therein is comparable to that in humans[7,8], the resulting SNP density of one SNP per 500 base pairs is considerably higher than is commonly used in human studies[6].

To evaluate the feasibility of GWA studies in this organism, we generated or assembled a variety of phenotypes. The phenotypes broadly fell into four categories: 23 were related to flowering under different environmental conditions; 23 were related to defence, ranging from recognition of specific bacterial strains to trichome density; 18 were element concentrations measured using inductively coupled plasma mass spectroscopy ('ionomics'); and 43 were loosely defined developmental traits, including dormancy and plant senescence. For details about each phenotype, see Supplementary Tables 2–5. The flowering phenotypes were generally strongly positively correlated, and were also negatively correlated with some other phenotypes, for example those related to size at flowering (Supplementary Fig. 9).

We first assessed evidence of association between each SNP and phenotype using the non-parametric Wilcoxon rank-sum test (Fisher's exact test was used for the small number of phenotypes that were categorical rather than quantitative). A significant difference between our study and the human GWA studies published so far is that our study population was heavily structured, and there is thus every reason to expect increased false-positive rates[9]. Indeed, most phenotypes gave rise to a distribution of *P* values that was strongly skewed towards zero (Supplementary Information, section 2.1.6). Figure 1a shows the number of distinct peaks of association identified for each phenotype using different *P*-value thresholds, as well as the number expected by chance alone. There is an excess of strong associations across phenotypes, as expected given the presence of confounding population structure (although we also expect some of these associations to be true). Furthermore, the degree of confounding varies greatly between phenotypes. Phenotypes related to flowering are generally more strongly affected, as would be expected given the correlation between flowering and geographic origins.

The population structure in our sample is highly complex, involving patterns of relatedness on all scales (Supplementary Fig. 4). As in previous studies[9], we found that, at least in terms of producing a *P*-value distribution that does not show obvious signs of confounding, statistical methods commonly used to control population structure in human genetics[10,11] fail to correct this confounding, whereas the mixed-model approach introduced by maize geneticists[12] seems to perform well (Supplementary Information, section 2.1.6). Figure 1b shows the number of peaks of association identified using this approach (as implemented using the program EMMA[13]). The excess of nominally significant association for flowering-related phenotypes

[1]Molecular and Computational Biology, [2]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA. [3]Department of Ecology & Evolution, University of Chicago, Chicago, Illinois 60637, USA. [4]Bindley Bioscience Center, [5]Purdue University, West Lafayette, Indiana 47907, USA. [6]Laboratoire de Génétique et Evolution des Populations Végétales, UMR CNRS 8016, Université des Sciences et Technologies de Lille 1, F-59655 Villeneuve d'Ascq Cedex, France. [7]Howard Hughes Medical Institute, La Jolla, California 92037, USA. [8]Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA. [9]Department of Cell and Development Biology, John Innes Centre, Norwich NR4 7UH, UK. [10]Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany. [11]Sainsbury Laboratory, Norwich NR4 7UH, UK. [12]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany. [13]Gregor Mendel Institute, A-1030 Vienna, Austria.
*These authors contributed equally to this work.

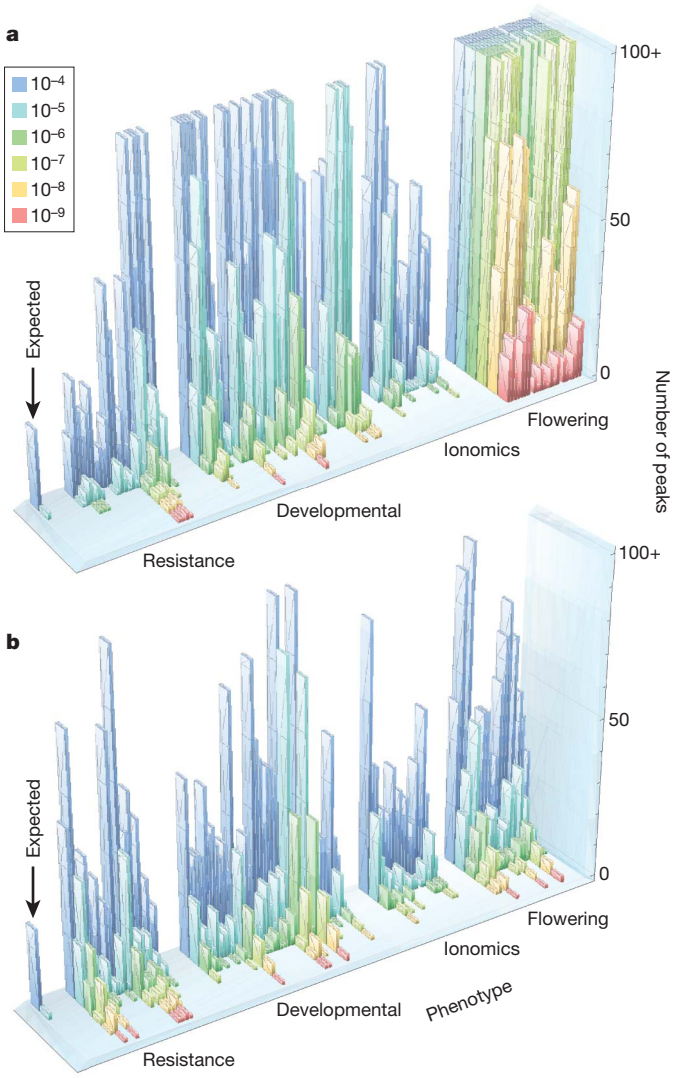

Figure 1 | The number of associations identified using different *P*-value thresholds for each phenotype. For each phenotype, the numbers of distinct peaks of association significant at nominal *P*-value thresholds (colour scale) are shown. The number of SNPs (out of 250,000) that would be expected to exceed each threshold is shown for comparison. **a**, No correction for population structure (non-parametric Wilcoxon rank-sum test). **b**, Correction for population structure (parametric mixed model (EMMA)).

has been eliminated, as would be expected if this excess were mostly due to confounding by population structure. There is a marked reduction in the number of associations of moderate significance (for example at $P \approx 10^{-4}$) across phenotypes, but the excess of highly significant associations persists (or has even become greater). This is precisely what would be expected from an increase in statistical power by switching to a parametric method that reduces confounding. It is tempting to conclude that most of these extreme *P* values must represent true associations, but there are reasons to be sceptical. First, although EMMA appears to produce a *P*-value distribution that conforms to the null expectation (except for extreme values; see Supplementary Figs 12–118), it seems almost certain that some confounding remains (see below). Second, simulation studies suggest that the *P* values produced by EMMA are not always well estimated and should be interpreted with caution (Supplementary Information, section 2.1.5).

It is thus not straightforward to distinguish true associations from spurious, regardless of whether we correct for population structure. There is no doubt, however, that there is real information in the data. Regardless of method used, six phenotypes yield single, strong peaks of association that can be seen by inspection. In all cases, the association results effectively identify single genes and correspond to known functional polymorphisms. An example of this is shown in Fig. 2, where the hypersensitive response to the bacterial avirulence gene *AvrRpm1* directly identifies the corresponding resistance gene *RESISTANCE TO P. SYRINGAE PV MACULICOLA 1* (*RPM1*)[14]. Similar results were obtained for other disease-resistance responses, sodium concentration, lesioning and *FRIGIDA* (*FRI*) expression (see Supplementary Figs 35–38, 60, 76 and 21, respectively).

More generally, SNPs closely linked to genes that are a-priori likely to be responsible for a particular phenotype are significantly over-represented among SNPs associated with that phenotype. For many of the phenotypes analysed, it is possible to predict, on the basis of existing functional knowledge, which genes might be important in natural variation. For these phenotypes, we determined which of our SNPs were located within 20 kilobases (kb) of an a-priori candidate, and tested whether these SNPs were over-represented among nominally significant associations. Figure 3 illustrates the procedure for the phenotype of flowering time at 10 °C. For example, SNPs with *P* values less than 10^{-3} from EMMA and *P* values less than 10^{-5} from the Wilcoxon rank-sum test are 2.7 times more likely to be close to candidate genes than are randomly chosen SNPs. This simultaneously demonstrates that background functional knowledge about flowering pathways helps predict which genes are involved in natural variation and that our GWA results identify many true associations. Indeed, by assuming that all associations not involving a-priori candidates are false, we see that the reciprocal of the enrichment ratio provides a crude upper bound for the false-discovery rate among the a-priori candidates (Supplementary Information, section 3.2). Continuing the example in Fig. 3, no more than 40% of the candidate SNPs that are significant using both tests are false. This is an upper bound because it seems almost certain that many of the (often much larger set of) strong associations that are not close to a-priori candidates will also turn out to be real.

As illustrated in Fig. 3, a-priori candidates are over-represented among strongly associated SNPs regardless of whether or not we

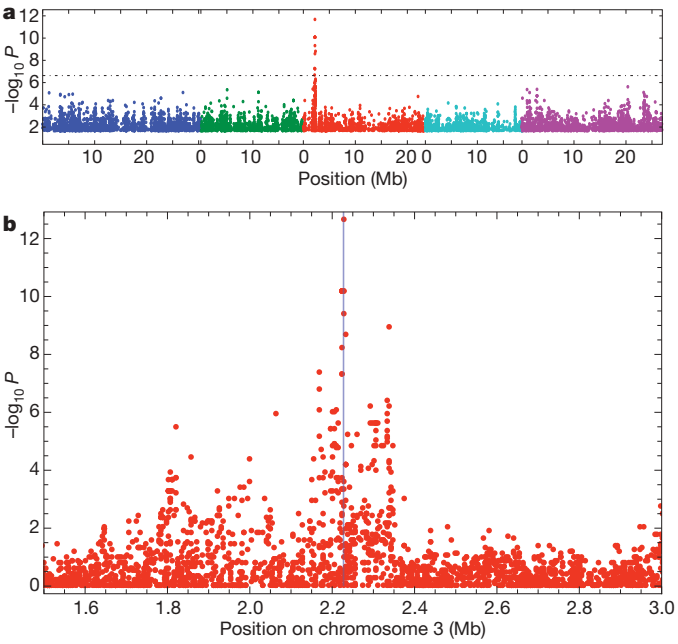

Figure 2 | GWA analysis of hypersensitive response to the bacterial elicitor *AvrRpm1*. **a**, Genome-wide *P* values from Fisher's exact test. The horizontal dash–dot line corresponds to a nominal 5% significance threshold with Bonferroni correction for 250,000 tests. **b**, Magnification of the genomic region surrounding *RPM1*, the position (and extent) of which is indicated by the vertical blue line. Mb, megabase.

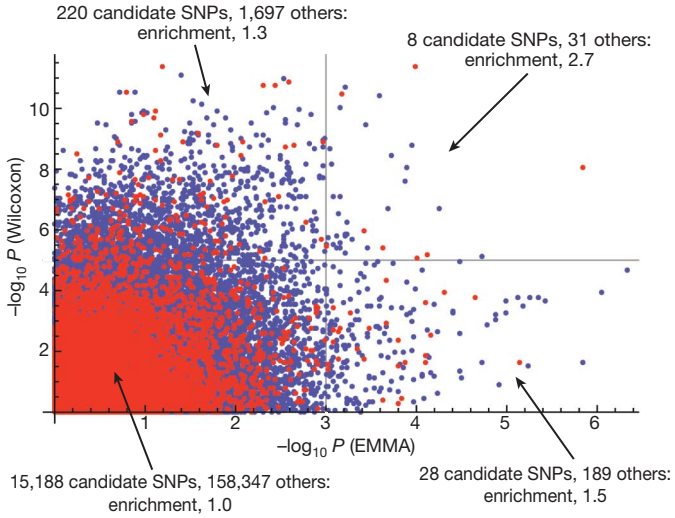

Figure 3 | Candidate SNPs are over-represented among strong associations. GWA analysis of the phenotype of flowering time at 10 °C: *P* values from the Wilcoxon rank-sum test are plotted against those from EMMA. Points corresponding to SNPs within 20 kb of a candidate gene are shown in red; the rest are shown in blue. The enrichment of the former over the latter in different parts of the distribution is as indicated.

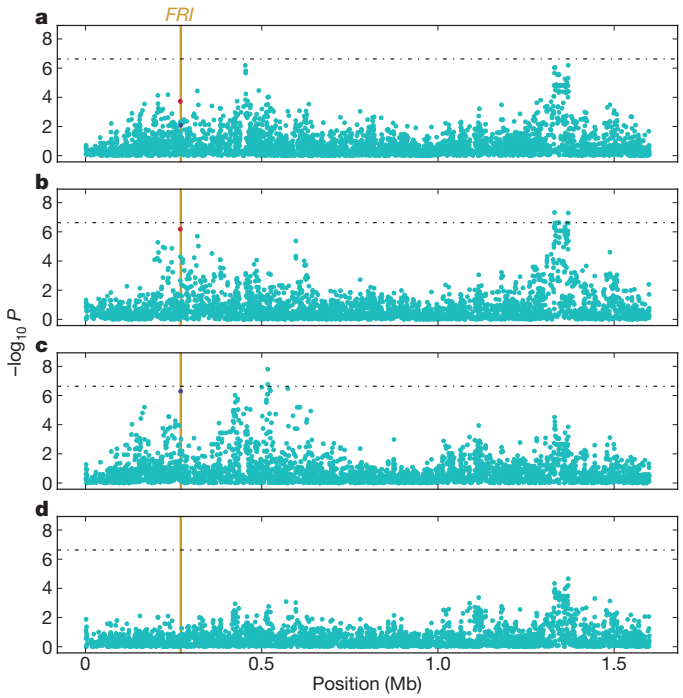

Figure 4 | Association with *FLC* expression at the top of chromosome 4 near *FRI*. The *P* values are from EMMA and the position of *FRI* is indicated by a vertical yellow line. The blue and red dots correspond to the Columbia and Landsberg *erecta* alleles of *FRI*, respectively. The horizontal dash–dot lines correspond to a nominal 5% significance threshold with Bonferroni correction for 250,000 tests. **a**, Single-SNP tests. **b**, Test with the Columbia allele included as a cofactor in the model. **c**, Test with the Landsberg *erecta* allele included as a cofactor in the model. **d**, Test with both alleles included as cofactors in the model.

correct for population structure, and the SNPs identified are not necessarily the same. Enrichment is greatest among SNPs that are strongly associated using both methods, however. This is true across the flowering-related phenotypes (Supplementary Fig. 10), and it is thus clear that both methods have utility. More stringent thresholds typically yield stronger enrichment but the variance also increases because the number of significant genes decreases, and there is thus no simple relationship between degree of enrichment and its statistical significance (Supplementary Fig. 11). Results for the other phenotypes are consistent with those for the flowering-related traits, but the candidate gene lists are too short for statistical analysis (Supplementary Information, section 3.1).

An additional problem in identifying true positives was the existence of complex peaks of association. Although many peaks were sharply defined and clearly identified a small number of genes (Fig. 2b), others were much more diffuse, sometimes covering several hundred kilobases without a clear centre. Figure 4 shows an example of such a peak and also suggests an explanation for their existence. The figure shows the pattern of association with *FLOWERING LOCUS C* (*FLC*) expression in the chromosomal region containing the vernalization-response gene *FRI*. Polymorphisms in *FRI* are known to affect flowering time partly through their effect on expression of *FLC*[5,15]. SNPs in the *FRI* region should thus be associated with *FLC* expression. This is indeed the case, but rather than a single, sharp peak of association centred on *FRI*, we have a 'mountain range' covering 500 kb and on the order of a hundred genes (Fig. 4a).

That *FRI* should be surrounded by a wide peak of association is, in itself, not unexpected given that the two common loss-of-function alleles at *FRI* appear to have been the subject of recent positive selection[16]. Indeed, the entire range collapses if these two alleles are added as cofactors to the model (Fig. 4d). More surprising is the fact that these two causal polymorphisms do not have the strongest association within the region. If we reduce allelic heterogeneity by factoring out one or other of the two alleles, the significance of the remaining polymorphism increases but it is still not the most significant in the region (Fig. 4b, c). A likely explanation for this is that some SNPs in the region are positively correlated (in linkage disequilibrium) with one of the *FRI* alleles (because of linkage) and the genomic background (because of population structure). This dual confounding is sufficient to make some of these SNPs more strongly associated with the phenotype than the true positives.

Given the difficulties described above, deciding which associations are worth following up must necessarily be highly subjective. The strongest associations do not always correspond to obvious candidates and are perhaps more interesting than associations in genes with known functions. However, in the absence of further evidence there is little point in discussing these associations. Supplementary Table 6 lists some of the most promising associations: additional a-posteriori candidates for each phenotype are given in Supplementary Figs 12–118. The genes listed were selected on the basis of annotation from within a 20-kb window surrounding each of the 500 most strongly associated SNPs (with a minor allele frequency of ≥0.1 for EMMA), distinguishing between those that had been considered candidates a priori and those that had not (those in the second category are marked with asterisks in the tables). As demonstrated in Fig. 3, we expect a high fraction of the associated a-priori candidates to be real. The full data are available through the project website (http://arabidopsis.usc.edu).

For the flowering-related phenotypes, one of the most striking findings was the strong correlation between phenotypes generated under very different growth-chamber and greenhouse conditions (Supplementary Table 2 and Supplementary Fig. 9; note that phenotypes from a field experiment were much less strongly correlated). As expected given the correlation in phenotypes, there are several regions of association that are shared across the majority of the flowering phenotypes (Supplementary Fig. 119). These regions vary considerably in width and many of them are complex in the sense of Fig. 4, perhaps as a consequence of strong selection. As expected given the results presented in Fig. 3, several of these regions coincided with a-priori candidates, such as *FRI*[15] and *FLC*[17] (Supplementary Table 6). Another interesting candidate is *DELAY OF GERMINATION 1* (*DOG1*)[18], which, though not originally thought to be involved with flowering, is highly associated with 20 different flowering phenotypes.

Among the other phenotypes, three previously identified disease-resistance-gene polymorphisms were readily identified[8], as were genes known to be involved with variation in sodium[19] and molybdenum[20] levels (Supplementary Table 6). Four genes known to be involved in trichome formation were strongly associated with both trichome phenotypes: one of these associations has recently been confirmed[21]. Finally, *ACCELERATED CELL DEATH 6* (*ACD6*), which has been experimentally shown to be directly involved in lesioning[22], is detected here as associated with several lesioning and chlorosis phenotypes. This association has also recently been experimentally confirmed (M.T. *et al*., unpublished observations).

By the standards of human GWA studies, the sample sizes used in this study (~96 or ~192 lines) are very low. It may thus seem surprising that we are able to map anything at all. However, power depends on the genetic architecture of the traits as well, and this works in our favour in at least two ways. First, we clearly find common alleles of major effect. Although effect sizes are hard to estimate for the same reason *P* values are, we note, for example, that in 44 phenotypes at least one of the 50 most strongly associated SNPs with a minor allele frequency greater than 15% explains more than 20% of the phenotypic variance (effect-size estimates and allele frequencies for every association are on the project website; see above). This is very different from human studies, which have generally identified only polymorphisms of very small phenotypic effect. The difference is probably due to the fact that, whereas human studies have focused on traits that are either deleterious or under strong stabilizing selection (for which a very strong trade-off between allelic effect and frequency is expected[23]), we are working with adaptively important variation. Indeed, human GWA studies focusing on traits such as skin colour seem to yield results more like those presented here[24,25].

Second, our study takes full advantage of the fact that we are working with inbred lines that can be grown in replicate under controlled conditions, making it possible to study multiple phenotypes while controlling environmental noise. Partly as a result of this, heritabilities for the traits studied are generally high, ranging from 42% for aphid number to over 99% for several flowering traits (Supplementary Table 7).

Without these two advantages, the amount of genotyping required would have made a study such as the present one prohibitively expensive. That said, there is little doubt that power in our study is severely limited by sample size. Simulation studies indicate that our power to detect alleles similar to those actually detected in the study is often no more than 30–40% using a sample size of 96 (results not shown). Increasing the sample size to 192 typically more than doubles power. Over 1,000 lines genotyped with our 250,000-SNP chip will soon be available: we look forward to seeing associations from these lines. Efforts are also under way to sequence the genomes of all these lines (http://www.1001genomes.org)—this will greatly facilitate follow-up studies, but we do not expect a massive increase in power as a result of this, because the SNP density used here seems adequate.

Power also depends on sample composition. In comparison with typical human GWA studies, our sample is characterized by having an extremely complex population structure. This is expected given that our global sample was collected partly to study population structure in *A. thaliana*[4]. GWA studies that use different samples (including regional, more homogeneous ones) are under way.

The fact that population structure can cause confounding and lead to an increased false-positive rate is well known, and the relative advantages of alternative statistical methods to correct for this have been much debated[10–12,26]. We feel that the discussion of this phenomenon has often been misleading, in that population structure is neither necessary nor sufficient for confounding to occur. At least for complex traits, the problem is better thought of as model misspecification: when we carry out GWA analysis using a single SNP at a time (as was done here and in most other previous GWA studies), we are in effect modelling a multifactorial trait as if it were due to a single locus. The polygenic background of the trait is ignored, as are other unobserved variables. This kind of marginal analysis is valid as long as the background is adequately captured by a variance term (or similar), but if the background variables are correlated with the SNP included in the model, bias will result. Population structure will lead to correlations (that is, linkage disequilibrium) between unlinked loci, and this will usually (but not always[27]) lead to confounding. Positive correlations are also expected as a result of strong selection. Both factors are likely to be important in the present study: for example, it is easy to imagine that plants from northern Sweden will tend to share cold-adaptive alleles at many causal loci as a result of selection, and marker alleles genome-wide as a result of demographic history.

This way of thinking about the problem helps us to interpret many of the results presented above. First, it becomes clear why GWA works so well for traits that are monogenic, or at least mostly due to a single, major locus. Examples in our study include the responses to disease-resistance genes (*RPM1*, *RESISTANCE TO P. SYRINGAE 2* (*RPS2*) and *RESISTANCE TO P. SYRINGAE 5* (*RPS5*)), *FRI* expression (*FRI* itself) and lesioning (*ACD6*). In all cases, GWA yields unambiguous results regardless of whether we correct for population structure. The reason is not that there is no confounding in these cases. The problem that has received so much attention in human genetics—inflated significance among unlinked, non-causal loci—is present, but with truly genome-wide coverage this is not very important because the true positive is expected to show the strongest association.

Second, it helps us explain the occurrence of broad, complex regions of association. As exemplified in Fig. 4, these can arise when SNPs in a region containing a major causative allele are positively correlated not only with that causative allele (owing to linkage disequilibrium in the narrow sense), but also with the genomic background (because of population structure and/or natural selection). The paradoxical consequence is that instead of a single peak of association centred on the causative locus, we expect a complex 'mountain landscape' in which many non-causal markers show stronger association than the causative allele itself. This makes it difficult to identify the causal variant within such regions. This type of confounding does not appear to have been recognized in the literature, and probably deserves more attention.

Third, it is clear that we should not expect statistical methods that are designed to take genome-wide patterns of relatedness into account[10–12] to correct for confounding that is due to selection generating correlations between causal loci. These types of methods will work only to the extent that the loci responsible for the genomic background have allele frequency distributions that are similar to those of non-causal loci, which is expected only if selection on each locus is weak. Accurate estimation of the size of the effects of the many candidate polymorphisms identified here (including distinguishing it from zero) will require either crosses or transgenic experiments. Any cross will eliminate long-range linkage disequilibrium, and short-range linkage disequilibrium can be overcome by choosing appropriate parental strains. For example, the *FRI* region in Fig. 4 also contains a promising association at *CRYPTIC PRECOCIOUS* (*CRP*), less than 100 kb away from *FRI* (Supplementary Fig. 127). If polymorphism at *FRI* is taken into account, *CRP* is no longer significantly associated, but this does not necessarily mean that the association is spurious. The two genes are too closely linked to be separated using standard crosses, but we can select parental strains that segregate for *CRP* but not *FRI*.

Such crosses are currently being carried out, by us and by other members of the *Arabidopsis* research community. We also anticipate that many more phenotypes will be generated and added to our public database. By combining results from GWA, linkage mapping and perhaps also intermediate phenotypes (for example expression data), it will be possible to make progress on deconstructing the regulatory networks that determine natural variation.

As genotyping and sequencing costs continue to decrease, GWA studies will become a standard tool for dissecting natural variation. It

is thus important to recognize their limitations. The problems raised here are not unique to *A. thaliana*. GWA alone will often not allow accurate estimate of allelic effects. It must also be remembered that all mapping studies are biased in the sense that they can only detect alleles that explain a sufficient fraction of the variation in the mapping population[28]. The present study can detect only alleles that are reasonably common in our global sample. A GWA study using a more local sample would undoubtedly uncover more variants that are locally common, and linkage mapping will identify major polymorphisms that happen to be segregating in the cross, even if one of the alleles is extremely rare in natural populations. The 'genetic architecture' of a trait depends on the population studied. To determine how it affects selection and evolution, we thus also need to understand the spatial and temporal scales on which selection is important.

METHODS SUMMARY

Because slightly different sets were used in different phenotyping experiments, the total number of lines used was 199 (Supplementary Table 1). Genotyping was done using standard protocols, and a combination of SNP calling and imputation algorithms were used to analyse the results (Supplementary Information, section 1). We called 216,130 SNPs, at an estimated error rate of 1.6%. GWA analysis was done with and without correction for confounding. For analysis with confounding, a mixed-model[12] implemented in the program EMMA[13] was used. For analysis without confounding, the Wilcoxon rank-sum test was used for ordinal data and Fisher's exact test was used for categorical data. Enrichment for candidate genes was investigated using lists of a-priori candidates identified from the literature.

Received 23 June; accepted 30 December 2009.
Published online 24 March 2010.

1. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6,** 95–108 (2005).
2. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
3. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in *Arabidopsis thaliana. Annu. Rev. Plant Biol.* **55,** 141–172 (2004).
4. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana. PLoS Biol.* **3,** e196 (2005).
5. Shindo, C. *et al.* Role of *FRIGIDA* and *FLC* in determining variation in flowering time of *Arabidopsis thaliana. Plant Physiol.* **138,** 1163–1173 (2005).
6. Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana. Nature Genet.* **39,** 1151–1155 (2007).
7. Nordborg, M. *et al.* The extent of linkage disequilibrium in *Arabidopsis thaliana. Nature Genet.* **30,** 190–193 (2002).
8. Aranzana, M. J. *et al.* Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* **1,** e60 (2005).
9. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3,** e4 (2007).
10. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67,** 170–181 (2000).
11. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38,** 904–909 (2006).
12. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38,** 203–208 (2005).
13. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178,** 1709–1723 (2008).
14. Grant, M. R. *et al.* Structure of the *Arabidopsis RPM1* gene enabling dual-specificity disease resistance. *Science* **269,** 843–846 (1995).
15. Johanson, U. *et al.* Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290,** 344–347 (2000).
16. Toomajian, C. *et al.* A non-parametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol.* **4,** e137 (2006).
17. Michaels, S. D. & Amasino, R. M. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* **11,** 949–956 (1999).
18. Bentsink, L., Jowett, J., Hanhart, C. J. & Koornneef, M. Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis. Proc. Natl Acad. Sci. USA* **103,** 17042–17047 (2006).
19. Rus, A. *et al.* Natural variants of At*HKT1* enhance Na^+ accumulation in two wild populations of *Arabidopsis. PLoS Genet.* **2,** e210 (2006).
20. Baxter, I. *et al.* Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (*MOT1*). *PLoS Genet.* **4,** e1000004 (2008).
21. Hilscher, J., Schlötterer, C. & Hauser, M.-T. A single amino acid replacement in ETC2 acts as major modifier of trichome patterning in natural *Arabidopsis* populations. *Curr. Biol.* **19,** 1747–1751 (2009).
22. Lu, H., Rate, D. N., Song, J. T. & Greenberg, J. T. *ACD6*, a novel ankyrin protein, is a regulator and an effector of salicylic acid signaling in the *Arabidopsis* defense response. *Plant Cell* **15,** 2408–2420 (2003).
23. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
24. Han, J. *et al.* A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **4,** e1000074 (2008).
25. Stokowski, R. P. *et al.* A genomewide association study of skin pigmentation in a South Asian population. *Am. J. Hum. Genet.* **81,** 1119–1132 (2007).
26. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
27. Rosenberg, N. A. & Nordborg, M. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed, or spatially distributed populations. *Genetics* **173,** 1665–1678 (2006).
28. Nordborg, M. & Weigel, D. Next-generation genetics in plants. *Nature* **456,** 720–723 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Carvalho for his advice on how to modify the OLIGO package. This work was primarily supported by US National Science Foundation (NSF) grant DEB-0519961 (J.B., M.N.), US National Institutes of Health (NIH) grant GM073822 (J.O.B.), and NSF grant DEB-0723935 (M.N.). Additional support was provided by the Dropkin Foundation, NIH grant GM057994 and NSF grant MCB-0603515 (J.B.), the Max Planck Society (D.W., M.T.), the Austrian Academy of Sciences (M.N.), the University of Lille 1 (F.R.), NIH grant GM078536 and NIH grant P42ES007373 (D.E.S.), NIH grant GM62932 (J.C., D.W.), the Howard Hughes Medical Institute (J.C.), the Deutsche Forschungsgemeinschaft (DFG) SFB 680 (J.d.M.), a Marie Curie International Outgoing Fellowship 'ANAVACO' (project number 220833; G.W.), and a Gottfried Wilhelm Leibniz Award of the DFG (D.W.). The project would not have been possible without the existence of The Arabidopsis Information Resource (http://arabidopsis.org).

Author Contributions J.O.B., J.B. and M.N. are equal senior authors. J.R.E. and D.W. generated the SNPs used in this project. S.A., M.H., Y.L., N.W.M., X.Z., J.O.B. and J.B. were responsible for the experimental aspects of genotyping. Y.S.H., B.J.V., M.H., T.T.H., R.J., X.Z., M.A.A., P.M., J.O.B., J.B. and M.N. were responsible for data management and the bioinformatics pipeline. S.A., I.B., B.B., J.C., C.D., M.D., J.d.M., N.F., J.M.K., J.D.G.J., T.M., A.N., F.R., D.E.S., C.T., M.T., M.B.T., D.W., J.B. and M.N. were responsible for phenotyping. S.A., Y.S.H., B.J.V., G.W., D.M., A.P., A.M.T., P.M. and M.N carried out the GWA analyses. Y.S.H. and D.M. developed the project website. M.N. wrote the paper with significant contributions from S.A., Y.S.H., B.J.V., G.W., A.P. and J.B. J.O.B., J.B. and M.N. designed and supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).