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

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

The model plant *Arabidopsis thaliana* is ideally suited for genome-wide association (GWA) studies in that it naturally occurs as inbred lines which can be genotyped once and phenotyped repeatedly. We demonstrate the power of this approach by carrying out a GWA study of 107 different phenotypes in a common set of up to 194 inbred lines genotyped for 216,130 SNPs. The results varied considerably between phenotypes. Some yielded unambiguous results in the form of distinct, obviously significant associations, usually corresponding to single genes, often *a priori* candidates. The majority of phenotypes yielded results that are harder to interpret because confounding by complex genetics and population structure make it difficult to distinguish true from false associations. However, *a priori* candidates are significantly overrepresented among these associations as well, making many of them excellent candidates for follow-up experiments by the *Arabidopsis* community. A public website has been developed to facilitate this. Our results are dramatically different from the results of human GWA studies in that we identify many common alleles with major effect. Our study clearly demonstrates the feasibility of GWA studies in *A. thaliana*, and suggests that the approach will be appropriate for many other organisms.

Although pioneered by human geneticists as a potential solution to the challenging problem of finding the genetic basis of common human diseases[1,2], advances in genotyping and sequencing technology have made GWA studies 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. This makes 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.

A. thaliana is predominantly selfing and naturally exists in the form of inbred lines, large numbers of which are available from stock centers. It occurs widely throughout the northern hemisphere and harbors considerable genetic variation for many adaptively important traits[3]. The extent of linkage disequilibrium (LD) has been shown to be appropriate for GWA studies[4,5], and a custom Affymetrix SNP-chip containing 250,000 SNPs has recently been developed[6].

In this paper we report the first results from using this chip. We have genotyped 199 lines for which many phenotypic measurements were available, and used these data to carry out what is effectively the first GWA study outside humans.

Genotypes and phenotypes

The genotyped sample includes a core set of 95 lines[7] for which a wide variety of phenotypes were available, plus a second set of 96 for which many phenotypes related to flowering were available[8]. Because slightly different sets were used in different 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 analyze the results (see Supplementary Section 1). We

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

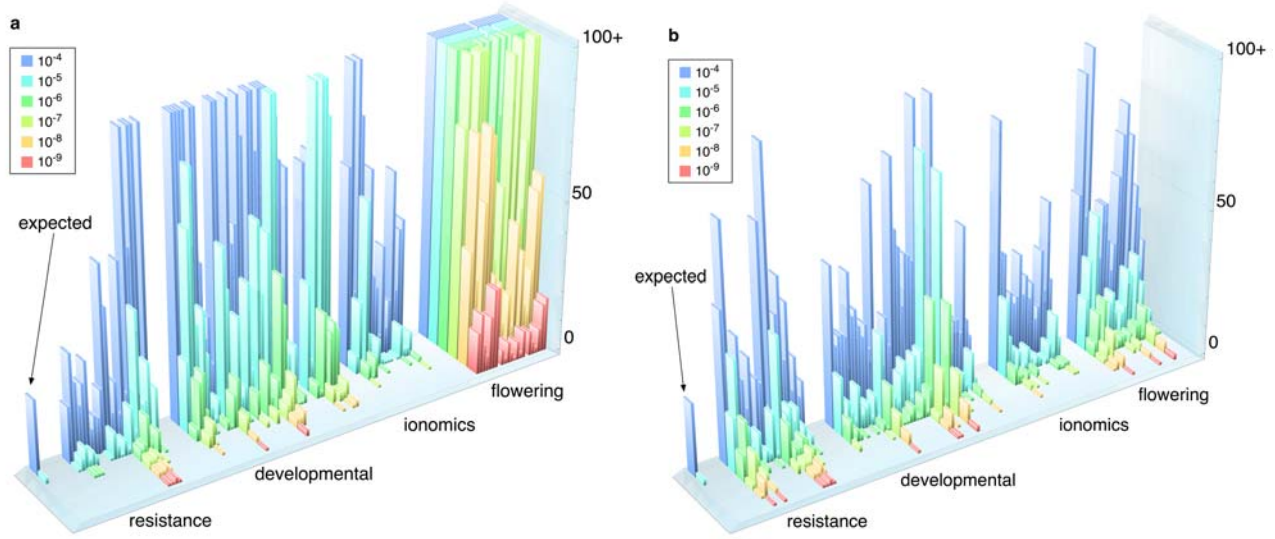

Figure 1 – The number of associations identified using different p-value thresholds for each phenotype. For each phenotype, the numbers of distinct peaks of association significant at nominal p-value thresholds of $10^{-4}, 10^{-5}, \ldots, 10^{-9}$ are shown. The number of SNPs (out of 250,000) that would be expected to exceed each threshold is shown for comparison. **a**, No correction for population structure (non-parametric Wilcoxon Test). **b**, Correction for population structure (parametric mixed model [EMMA]).

called 216,130 SNPs, at an estimated error rate of 1.6%. Since the genome of *A. thaliana* is around 120 million bp and the extent of LD comparable to that in humans, the resulting SNP density of one SNP per 500 bp is considerably higher than is commonly used in human studies[6].

To evaluate the feasibility of GWA studies in this organism, a variety of phenotypes were generated or assembled. The phenotypes broadly fell into four categories: 23 were related to flowering under different environmental conditions; 23 were related to defense, 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, see Supplementary Tables 2–5.

Analysis

Population structure and p-values. We first assessed evidence of association between each SNP and phenotype using the non-parametric Wilcoxon Test (Fisher's Exact Test was used for the small number of phenotypes that were categorical rather than quantitative). A major difference between our study and the human GWA studies published to date is that our study population is heavily structured, and there is thus every reason to expect elevated false-positive rates[9]. Indeed, most phenotypes gave rise to a distribution of p-values that was strongly skewed toward zero (Supplementary Section 2.1.4). Fig. 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 (see 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 for population structure in human genetics[10,11] fail to correct for population structure, whereas the mixed-model approach introduced by maize geneticists[12] appears to perform well (see Supplementary Section 2.1.4). Fig. 1b shows the number of peaks of association identified using this approach (as implemented in the program EMMA[13]). The excess of nominally significant association for flowering-related phenotypes 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 (e.g., p-values below 10^{-4}) across phenotypes, but the excess of highly significant associations clearly 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 skeptical.

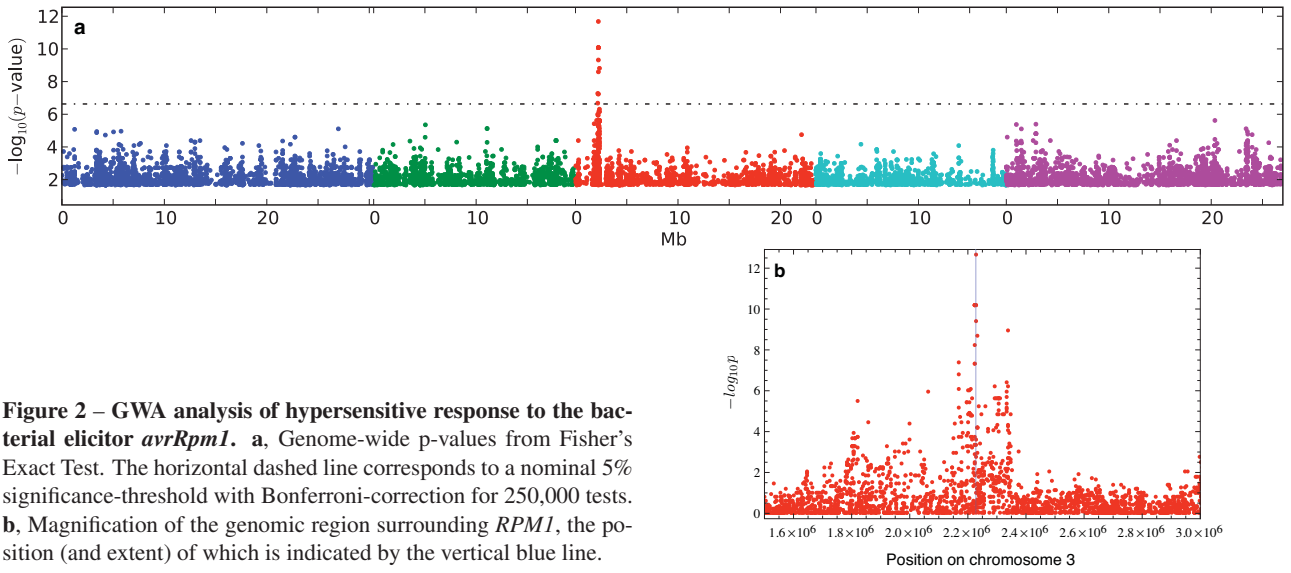

Figure 2 – GWA analysis of hypersensitive response to the bacterial elicitor *avrRpm1*. **a**, Genome-wide p-values from Fisher's Exact Test. The horizontal dashed line corresponds to a nominal 5% significance-threshold with Bonferroni-correction for 250,000 tests. **b**, Magnification of the genomic region surrounding *RPM1*, the position (and extent) of which is indicated by the vertical blue line.

First, although EMMA appears to produce a p-value distribution that conforms to the null-expectation (except for extreme values: see Supplementary Figs. 9–115), it seems almost certain that some confounding remains (see Discussion). Second, simulation studies suggest that the p-values produced by EMMA are not always well estimated and should be interpreted with caution (see Supplementary Section 2.1.5).

It is thus not straightforward to distinguish true from spurious associations, regardless of whether we correct for population structure. There is no doubt, however, that there is real signal in the data. Indeed, a small number of phenotypes yield single, strong peaks of association, effectively identifying single genes, regardless of method used. 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 (*R*) responses, sodium concentration, lesioning, and *FRIGIDA* (*FRI*)[15] expression (see below and Supplementary Figs. 32, 34, 57, 73, and 18, 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 analyzed it is possible to predict which genes might be important in natural variation based on existing functional knowledge. For these phenotypes we determined which of our SNPs were located within 20 kb of an *a priori* candidate, and tested whether these SNPs were over-represented among nominally significant associations. Fig. 3a illustrates the procedure for flowering time at 10°C (FT10). For example, SNPs with a p-value less than 10^{-3} from EMMA and a p-value less than 10^{-5} from the Wilcoxon 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 inverse of the enrichment ratio provides a crude upper bound for the false-positive rate among the *a priori* candidates (see Supplementary Section 3.2). Continuing the example in Fig. 3a, 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 (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. 3a, *a priori* candidates are over-represented among strongly associated SNPs regardless of whether we correct for population structure or not, and the SNPs identified are not necessarily the same. Enrichment is clearly greatest among SNPs that are strongly associated using both methods, however. This is true across the flowering-related phenotypes (Fig. 3b), and it is thus clear that both methods have utility. 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 Section 3.1).

Complex peaks. An additional problem in identifying true positives was the existence of complex peaks of association. While many peaks were sharply defined and clearly identified a small number of genes (illustrated in Fig. 2b), others were much more diffuse, sometimes covering several hundred kb without a clear center. Fig. 4 shows an example

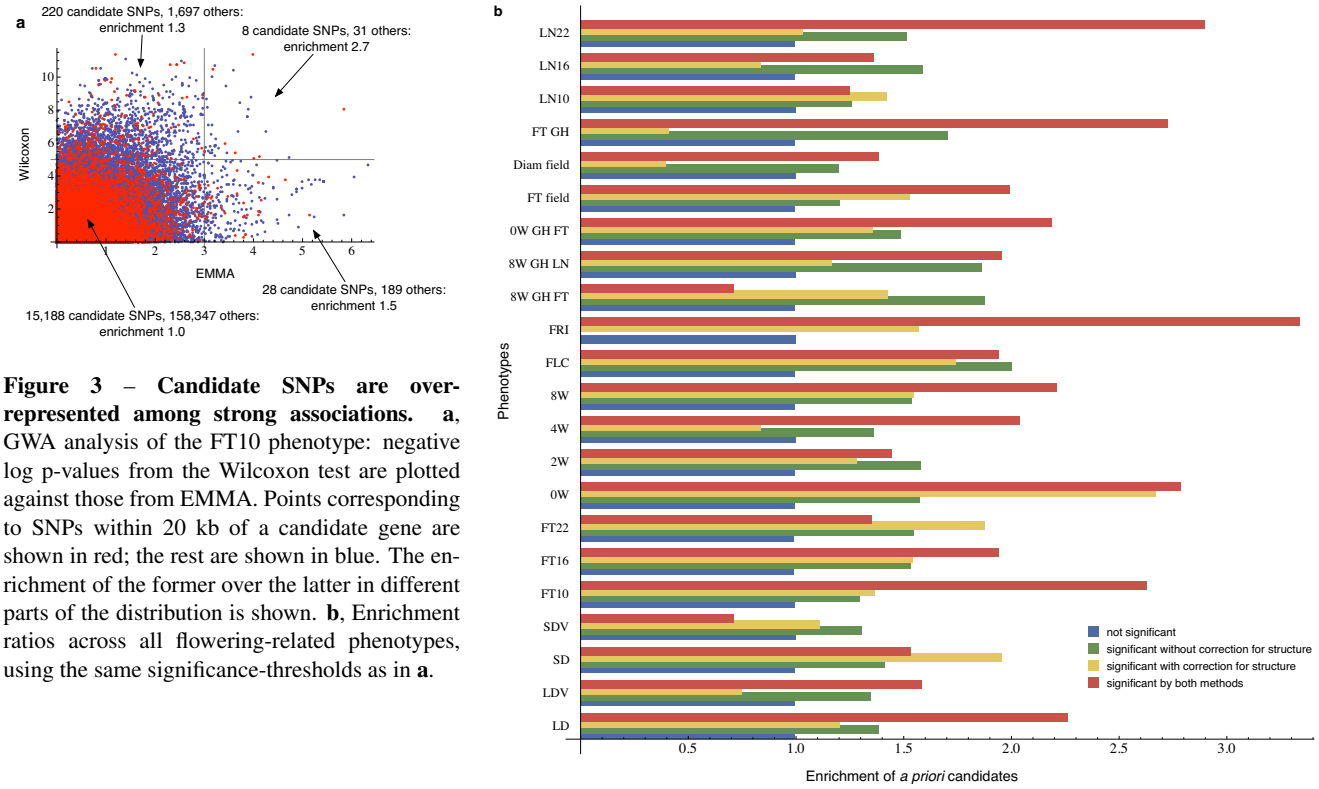

Figure 3 – Candidate SNPs are over-represented among strong associations. **a**, GWA analysis of the FT10 phenotype: negative log p-values from the Wilcoxon test are plotted against those from EMMA. Points corresponding to SNPs within 20 kb of a candidate gene are shown in red; the rest are shown in blue. The enrichment of the former over the latter in different parts of the distribution is shown. **b**, Enrichment ratios across all flowering-related phenotypes, using the same significance-thresholds as in **a**.

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*[8,15]. SNPs in the *FRI* region should thus be associated with *FLC* expression. This is indeed the case, but rather than a single peak of association centered 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 surprising 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 co-factors to the regression 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 the 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.

Associations

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 function. However, in the absence of further evidence there is little point in discussing these associations. The genes listed below were selected based on annotation from within a 20 kb window surrounding each of the top 500 most strongly associated SNPs (with minor allele frequency ≥ 0.1 for EMMA), distinguishing between those that had been considered candidates *a priori* and those that had not (the latter category is 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. Table 1 lists some of the most promising associations: additional *a posteriori* candidates for each phenotype are given in Supplementary Figs. 9–115. The full data are available through the project website (http://arabidopsis.usc.edu).

Flowering-related phenotypes. One of the most striking findings was the very high degree of concordance between the results for flowering phenotypes generated under differ-

Table 1 Selected candidates and phenotypic associations.

Gene name	Number of phenotypes			Nature of phenotype(s)	Suppl. Fig(s)
	Wilc.	EMMA	Both		
ARABIDOPSIS THALIANA HOMEOBOX 1 (*ATH1*)	18	12	9	Flowering	117
SHORT VEGETATIVE PHASE (*SVP*)	16	18	16	Flowering	118
FRIGIDA (*FRI*)	13	21	13	Flowering, life cycle and seedling dev.	119, 124
FLOWERING PROMOTING FACTOR 1 (*FPF1*)	6	7	4	Flowering	120
GIBBERELLIN 20-OXIDASE 3 (*YAP169*)	11	10	6	Flowering	121
RELATED TO AB13/VP1 (*RAV-1*)	11	6	5	Flowering	122
SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 4 (*SPL4*)	9	10	7	Flowering	123
CRYPTIC PRECOCIOUS (*CRP*)	8	14	6	Flowering	124
SPA1-RELATED 2 (*SPA2*)	16	7	6	Flowering	125
SPA1-RELATED 4 (*SPA4*)	10	10	7	Flowering	126
SIMILAR TO REDUCED VERNALIZATION RESPONSE (*sim to VRN1*)	18	13	11	Flowering	127
DELAY OF GERMINATION 1 (*DOG1*)	25	24	21	Flowering, life cycle, seedling dev.	128
DWARF IN LIGHT 2 (*DFL2*)	25	18	15	Flowering and seedling dev.	129
FLOWERING LOCUS C (*FLC*)	15	18	12	Flowering, life cycle, seedling dev.	130
ENHANCER OF TRY AND CPC 3 (*ETC3*)	16	18	14	Flowering	124
FLOWERING LOCUS T (*FT*)	2	11	1	Flowering, life cycle and seedling dev.	131
ENHANCER OF AG-4 2 (*HUA2*)	2	1	0	Flowering	132
HUA ENHANCER 2 (*HEN2*)	9	9	6	Flowering	133
VERNALIZATION INSENSITIVE 3 (*VIN3*)	7	5	3	Flowering	134
RESISTANCE TO P. SYRINGAE PV MACULICOLA 1 (*RPM1*)	3	4	3	Bacterial and fungal resistance	
CONSTITUTIVE TRIPLE RESPONSE (*CTR1*)	5	5	5	Bacterial and fungal resistance	135
RESISTANT TO P. SYRINGAE 2 (*RPS2*)	1	1	1	*avrRpt2*	136
RESISTANT TO P. SYRINGAE 5 (*RPS5*)	2	2	1	Bacterial resistance	137
TRICHOMELESS 1 and *2* (*TCL1*, *TCL2*) and *ENHANCER OF TRY AND CPC 2* (*ETC2*)	2	2	2	Trichomes	138
HIGH-AFFINITY K+ TRANSPORTER 1 (*HKT1*)	1	1	1	Sodium	140
MOLYBDATE TRANSPORTER 1 (*MOT1*)	1	0	0	Molybdenum	141
SULFATE TRANSPORTER 1;2 (*SULTR1;2*)	1	2	1	Sulfur and Selenium	142
COMATOSE (*CTS*)	6	4	4	Seedling development and flowering	143
LIGHT-DEPENDENT SHORT HYPOCOTYLS 1 (*LSH1*)	1	1	1	Seedling development	144
LONG HYPOCOTYL IN FAR-RED (*HFR1*)	3	1	0	Seedling development and rosette erect	145
ACCELERATED CELL DEATH 6 (*ACD6*)	12	14	11	Bacterial and fungal resistance, aphid number, trichomes, weight, lesioning and chlorosis	146
SUPPRESSOR OF G2 ALLELE OF SKP1 (*SGT1A*)	9	8	6	Bacterial and fungal resistance, trichomes, weight, lesioning	147
ARABIDOPSIS THALIANA HOMOLOG OF YEAST AUTOPHAGY 18 (*AtATG18h*)	3	4	3	weight, aphid number and lesioning	148
FLAVANOL SYNTHASE (*FLS*)	1	1	1	Anthocyanin presence at 22°C	149

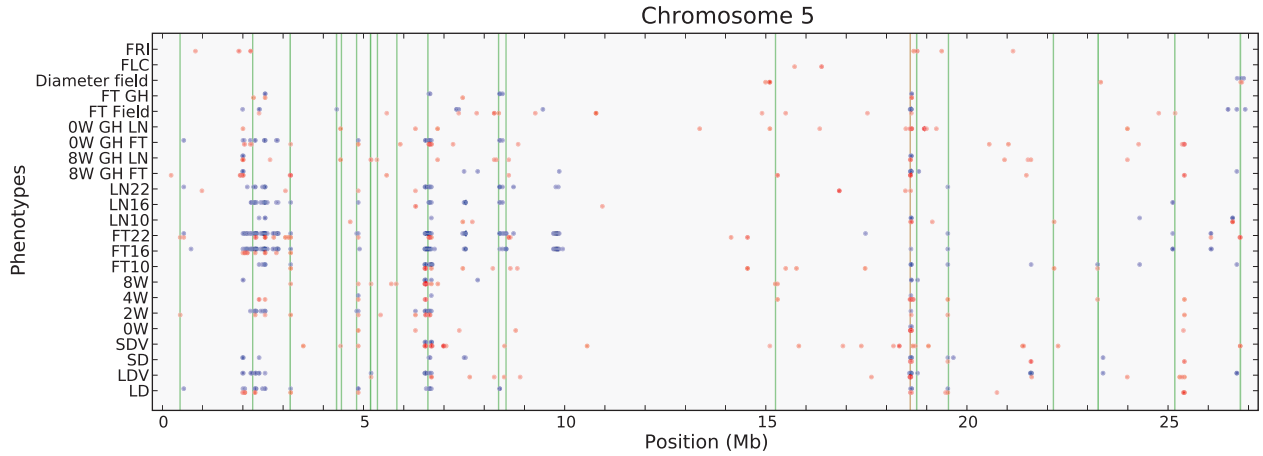

Figure 5 – The location of strong flowering-related associations on chromosome 5. Blue dots indicate SNPs significant at the 10^{-8} level using the Wilcoxon test; red dots indicate SNPs significant at the 10^{-4} level using EMMA. At most the top 0.1% of SNPs for each phenotype are plotted. Deeper color indicates stronger association. The green lines indicate the positions of *a priori* candidates within 20 kb of one of the dots. The orange line marks *DOG1*, a strong *a posteriori* candidate. For all chromosomes, see Supplementary Fig. 116.

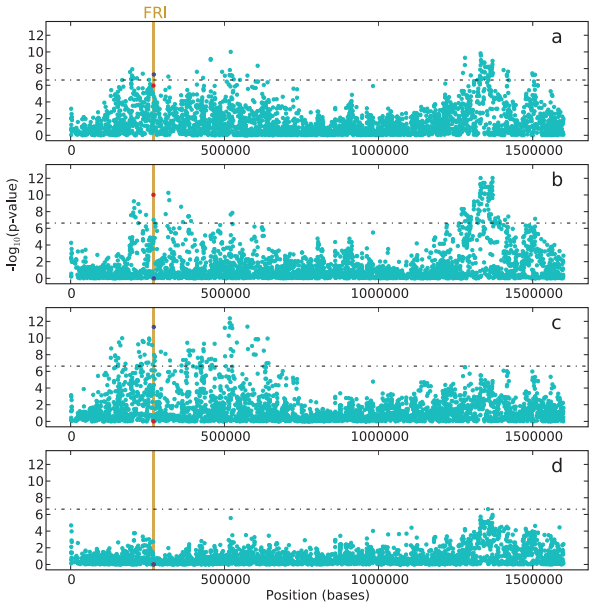

Figure 4 – Association with *FLC* expression at the top of chromosome 4 near *FRI*. The p-values are from simple linear regression on the SNPs; the position of *FRI* is indicated by a vertical yellow line. **a**, Regression only on the SNPs. **b**, Col-allele of *FRI* (blue dot) is added as co-factor in the regression model. **c**, Ler-allele of *FRI* (red dot) is added as co-factor in the regression model. **d**, Both alleles added as co-factors in the regression model.

ent environmental conditions (Supplementary Table 2). As illustrated in Fig. 5, there are several regions of association that are shared across the majority of the flowering phenotypes. The figure also illustrates the variability in the width of these regions. As expected given the results presented in Fig. 3, several of these regions coincided with *a priori* candidates. Particularly noteworthy in this respect were *ARABIDOPSIS THALIANA HOMEOBOX 1* (*ATH1*)[17], *SHORT VEGETATIVE PHASE* (*SVP*)[18], *FRI*[15], and *FLC*[19], which are associated with the majority of flowering-related phenotypes, regardless of test used. Similar patterns of association were shown by *SIMILAR TO VRN1*, which is a good candidate given its sequence similarity to *REDUCED VERNALIZATION RESPONSE* (*VRN1*)[20], *DELAY OF GERMINATION 1* (*DOG1*)[21], which, though not originally thought to be involved with flowering, is highly associated with 20 different flowering phenotypes, and the hypocotyl elongation gene *DWARF IN LIGHT 2* (*DFL2*)[22].

Defense phenotypes. Inoculation with *Pseudomonas syringae* pv. *syringae* DC3000 (DC3000) modified to express the *avr* genes *avrRpm1*/*avrB*, *avrRpt2* and *avrRps5* all produced impressive peaks in the previously identified cognate *R* genes *RPM1*[14], *RESISTANT TO P. SYRINGAE 2* (*RPS2*)[23] and *5* (*RPS5*)[24], respectively. *CONSTITUTIVE TRIPLE RESPONSE* (*CTR1*)[25], a gene involved in signaling by the hormone ethylene, known to be important for pathogen resistance, also resides in a smaller peak on chromosome 5 for *avrRpm1* and *avrB*.

Four genes known to be involved trichome-formation were found for both trichome phenotypes: *TRICHOMELESS 1* and *2* (*TCL1* and *TCL2*)[26], and *ENHANCER OF TRY AND CPC 2* (*ETC2*), all found in a single tandem repeat, and *TRANSPAR-*

ENT TESTA GLABRA 1 (*TTG1*)[27]. In addition, *ACCELERATED CELL DEATH 6* (*ACD6*)[28], which is involved in both defense and growth phenotypes, was also found (see below).

Ionomics phenotypes. The sodium (Na) association of *HIGH-AFFINITY K+ TRANSPORTER 1* (*HKT1*)[29] was readily confirmed. In addition, a good candidate for sulfur (S) was identified in the form of *SULFATE TRANSPORTER 1;2* (*SULTR1;2*)[30,31]; this gene also showed association with selenium (Se), which is significant because S and Se are chemical analogues and mutation of *SULTR1;2* causes reduced Se and S uptake[30]. Finally, the Wilcoxon test identified *MOLYBDATE TRANSPORTER 1* (*MOT1*[32]) for molybdenum.

Developmental phenotypes. Plausible candidates for germination-related phenotypes were found in the form of *COMATOSE* (*CTS*) which plays a key role in dormancy and germination[33] and both *LIGHT-DEPENDENT SHORT HYPOCOTYLS 1* (*LSH1*)[34] and *LONG HYPOCOTYL IN FAR-RED HFR1* (*HFR1*)[35] for hypocotyl elongation.

ACD6 has been experimentally shown to be directly involved in lesioning[28]. It is detected here as associated with several lesioning and chlorosis phenotypes. This association has been experimentally confirmed as being causal, as described in a companion paper[36]. *SUPPRESSOR OF G2 ALLELE OF SKP1* (*SGT1A*)[37], which regulates resistance gene triggered immunity, and *AtATG18h*, likely required for autophagasome formation and expressed during senescence or environmental stress[38], also make good candidates for the lesioning phenotypes.

Discussion

We have carried out a GWA study of 107 diverse phenotypes in a global sample of *A. thaliana* inbred lines and identified a very large number candidates for experimental verification. These studies will be facilitated by our public website. While our study is analogous to human efforts[1,2], there are major differences in both design and objectives. Here we contrast our study and results with those of human studies, and discuss the implications for the future of GWA studies, which will soon be routine in many organisms.

By the standards of human GWA studies, the sample sizes used here ($\sim$96 or $\sim$192 lines) are tiny. 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 favor in at least two ways. First, we clearly find common alleles of major effect. This is very different from human studies, which have generally identified only polymorphisms of very small phenotypic effect. The difference is likely 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[39,40]), we are working with adaptively important traits. We predict that human GWA studies focusing on traits like skin color would yield results more like those presented here.

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 is possible to study multiple phenotypes while controlling environmental noise. Partially as a result of this, heritabilities for the traits studied are high, ranging from around 60% for some of the ionomics traits to over 99% for several flowering traits.

Without these advantages, the amount of genotyping required would have made a study like 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[41]. Increasing the sample size to 192 typically more than doubles power[41]. Well over 1,000 lines genotyped with our 250k SNP chip will soon be available: we look forward to seeing associations from these lines. Efforts are also underway 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[41].

Power also depends on sample composition. In comparison with typical human GWA studies, our sample is characterized by extremely strong population structure. This is expected given that our global sample was collected partly to study population structure in *A. thaliana*[7]. GWA studies that utilize different samples (including regional, more homogeneous ones) are under way.

The fact that population structure can cause confounding and lead to an elevated false-positive rate is well known and the relative advantages of alternative statistical methods to correct for this have been much debated[10,11,12,42]. 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 mis-specification: when we carry out GWA analysis using a single SNP at a time (as is done in this, and most other GWA studies to date), we are in effect modeling a multi-factorial 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 causes no problem 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 (*i.e.*, linkage disequilibrium) between unlinked loci, and this will usually (but not always[43]) 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 interpret many of the results presented above. First, it becomes clear why GWA works so well for traits that are monogenic, or at least are mostly due to a single major locus. Examples in our study include the various *R* gene responses (*RPM1*, *RPS2*, and *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 clearly present, but with truly genome-wide coverage this does not really matter, because the true positive will 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 (due 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 centered on the causative locus, we expect a complex mountain landscape where many non-causal markers will 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[10,11,12] 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 very promising association at *CRP*, less than 100 kb away from *FRI* (see Supplementary Fig. 124). 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* 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 (e.g., expression data), it will be possible to make progress on deconstruct 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[44]. The present study can only detect 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. In order to determine how genetic architecture affects selection and evolution, we thus also need to understand the spatial and temporal scales over which selection is important.

REFERENCES

1. Hirschhorn, J. N. and Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat 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*, 447:661–78, 2007.

3. Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.*, 55:141–172, 2004.

4. Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A., and Weigel, D. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.*, 30: 190–193, 2002.

5. Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., and Nordborg, M. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.*, 1:e60, 2005.

6. Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., and Nordborg, M. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.*, 39:1151–1155, 2007.

7. Nordborg, M., Hu, T. T., 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, N. A., Shah, C., Wall, J., Wang, J., Zhao, K., Kalbfleisch, T., Schultz, V., Kreitman, M., and Bergelson, J. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.*, 3:e196, 2005.

8. Shindo, C., Aranzana, M. J., Lister, C., Baxter, C., Nicholls, C., Nordborg, M., and Dean, C. Role of *FRIGIDA* and *FLC* in determining variation in flowering time of *Arabidopsis thaliana*. *Plant Physiol.*, 138:1163–1173, 2005.

9. Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., and Nordborg, M. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.*, 3:e4, 2007.

10. Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. Association mapping in structured populations. *Am J Hum Genet*, 67:170–181, 2000.

11. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–9, 2006.

12. Yu, J., Pressoir, G., Briggs, W., Vroh Bi, I., Yamasaki, M., Doebley, J., McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., and Buckler, E. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38:203–208, 2006.

13. Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–23, 2008.

14. Grant, M. R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R. W., and Dangl, J. L. Structure of the *Arabidopsis Rpm1* gene enabling dual-specificity disease resistance. *Science*, 269:843–846, 1995.

15. Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., and Dean, C. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science*, 290:344–347, 2000.

16. Toomajian, C., Hu, T. T., Aranzana, M. J., Lister, C., Tang, C., Zheng, H., Calabrese, P., Dean, C., and Nordborg, M. A non-parametric test reveals selection for rapid flowering in the *Arabidopsis* genome. *PLoS Biol.*, 4:e137, 2006.

17. Proveniers, M., Rutjens, B., Brand, M., and Smeekens, S. The *Arabidopsis* TALE homeobox gene *ATH1* controls floral competency through positive regulation of *FLC*. *Plant Journal*, 52: 899–913, 2007.

18. Lee, J. H., Yoo, S. J., Park, S. H., Hwang, I., Lee, J. S., and Ahn, J. H. Role of *SVP* in the control of flowering time by ambient temperature in *Arabidopsis*. *Genes & Development*, 21:397–402, 2007.

19. Michaels, S. D. and 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.

20. Levy, Y., Mesnage, S., Mylne, J., Gendall, A., and Dean, C. Multiple roles of *Arabidopsis VRN1* in vernalization and flowering time control. *Science*, 297:243–6, 2002.

21. Bentsink, L., Jowett, J., Hanhart, C. J., and Koornneef, M. Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA*, 103(45): 17042–17047, 2006.

22. Takase, T., Nakazawa, M., Ishikawa, A., Manabe, K., and Matsui, M. *DFL2*, a new member of the *Arabidopsis GH3* gene family, is involved in red light-specific hypocotyl elongation. *Plant and Cell Physiology*, 44:1071–1080, 2003.

23. Bent, A. F., Kunkel, B. N., Dahlbeck, D., Brown, K. L., Schmidt, R., Giraudat, J., Leung, J., and Staskawicz, B. J. *Rps2* of *Arabidopsis thaliana* — A leucine-rich repeat class of plant-disease resistance genes. *Science*, 265:1856–1860, 1994.

24. Warren, R. F., Henk, A., Mowery, P., Holub, E., and Innes, R. W. A mutation within the leucine-rich repeat domain of the *Arabidopsis* disease resistance gene *RPS5* partially suppresses multiple bacterial and downy mildew resistance genes. *Plant Cell*, 10:1439–1452, 1998.

25. Lin, Z. F., Alexander, L., Hackett, R., and Grierson, D. *LeCTR2*, a *CTR1*-like protein kinase from tomato, plays a role in ethylene signalling, development and defence. *Plant Journal*, 54:1083–1093, 2008.

26. Wang, S. C., Kwak, S. H., Zeng, Q. N., Ellis, B. E., Chen, X. Y., Schiefelbein, J., and Chen, J. G. *TRICHOMELESS1* regulates trichome patterning by suppressing *GLABRA1* in *Arabidopsis*. *Development*, 134:3873–3882, 2007.

27. Bouyer, D., Geier, F., Kragler, F., Schnittger, A., Pesch, M., Wester, K., Balkunde, R., Timmer, J., Fleck, C., and Hulskamp, M. Two-dimensional patterning by a trapping/depletion mechanism: The role of *TTG1* and *GL3* in *Arabidopsis* trichome formation. *Plos Biology*, 6:1166–1177, 2008.

28. Lu, H., Rate, D. N., Song, J. T., and 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.

29. Rus, A., Baxter, I., Muthukumar, B., Gustin, J., Lahner, B., Yakubova, E., and Salt, D. E. Natural variants of *AtHKT1* enhance Na+ accumulation in two wild populations of *Arabidopsis*. *Plos Genetics*, 2:1964–1973, 2006.

30. Shibagaki, N., Rose, A., McDermott, J. P., Fujiwara, T., Hayashi, H., Yoneyama, T., and Davies, J. P. Selenate-resistant mutants of *Arabidopsis thaliana* identify *Sultr1;2*, a sulfate transporter required for efficient transport of sulfate into roots. *Plant Journal*, 29:475–486, 2002.

31. Yoshimoto, N., Takahashi, H., Smith, F. W., Yamaya, T., and Saito, K. Two distinct high-affinity sulfate transporters with different inducibilities mediate uptake of sulfate in *Arabidopsis* roots. *Plant Journal*, 29:465–473, 2002.

32. Baxter, I., Muthukumar, B., Park, H. C., Buchner, P., Lahner, B., Danku, J., Zhao, K., Lee, J., Hawkesford, M. J., Guerinot, M. L., and Salt, D. E. Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (*MOT1*). *Plos Genetics*, 4, 2008.

33. Carrera, E., Holman, T., Medhurst, A., Peer, W., Schmuths, H., Footitt, S., Theodoulou, F. L., and Holdsworth, M. J. Gene expression profiling reveals defined functions of the ATP-binding cassette transporter *COMATOSE* late in phase II of germination. *Plant Physiology*, 143:1669–1679, 2007.

34. Zhao, L., Nakazawa, M., Takase, T., Manabe, K., Kobayashi, M., Seki, M., Shinozaki, K., and Matsui, M. Overexpression of *LSH1*, a member of an uncharacterised gene family, causes enhanced light regulation of seedling development. *Plant Journal*, 37:694–706, 2004.

35. Kim, Y. M., Woo, J. C., Song, P. S., and Soh, M. S. *HFR1*, a phytochrome A-signalling component, acts in a separate pathway from *HY5*, downstream of *COP1* in *Arabidopsis thaliana*. *Plant Journal*, 30:711–719, 2002.

36. Todesco, M., Balasubramanian, S., Hu, T. T., Epple, P., Kuhns, C., Sureshkumar, S., Schwartz, C., Bomblies, K., Lanz, C., Laitinen, R. A. E., Chory, J., Lipka, V., Dangl, J. L., Nordborg, M., and Weigel, D. A fitness trade off between growth and disease resistance in *Arabidopsis thaliana*. Submitted to *Nature*, 2009.

37. Noel, L. D., Cagna, G., Stuttmann, J., Wirthmuller, L., Betsuyaku, S., Witte, C. P., Bhat, R., Pochon, N., Colby, T., and Parker, J. E. Interaction between *SGT1* and Cytosolic/Nuclear *HSC70* chaperones regulates *Arabidopsis* immune responses. *Plant Cell*, 19:4061–4076, 2007.

38. Xiong, Y., Contento, A. L., and Bassham, D. C. *AtATG18a* is required for the formation of autophagosomes during nutrient stress and senescence in *Arabidopsis thaliana*. *Plant Journal*, 42:535–546, 2005.

39. Hirschhorn, J. N. Genomewide association studies–illuminating biologic pathways. *The New England Journal of Medicine*, 360: 1699–701, 2009.

40. Goldstein, D. B. Common genetic variation and human traits. *The New England Journal of Medicine*, 360(17):1696–8, 2009.

41. Meng, D., Huang, Y., Vilhjálmsson, B., Platt, A., Borevitz, J. O., Bergelson, J., and Nordborg, M. The effect of SNP density and sample size on the power of genome-wide association studies in *Arabidopsis thaliana*. Submitted to *PLoS One*, 2009.

42. Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

43. Rosenberg, N. and 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.

44. Nordborg, M. and Weigel, D. Next-generation genetics in plants. *Nature*, 456:720–723, 2008.

45. Wolyn, D. J., Borevitz, J. O., Loudet, O., Schwartz, C., Maloof, J., Ecker, J. R., Berry, C. C., and Chory, J. Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. *Genetics*, 167:907–917, 2004.

46. Carvalho, B., Bengtsson, H., Speed, T. P., and Irizarry, R. A. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8:485–499, 2007.

47. Roberts, A., McMillan, L., Wang, W., Parker, J., Rusyn, I., and Threadgill, D. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics*, 23:i401–7, 2007.

48. Clark, R. M., Schweikert, G., Ossowski, S., Zeller, G., Toomajian, C., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R., and Weigel, D. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, 317:338–342, 2007.

49. Scheet, P. and Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–44, 2006.

50. Reignault, P., Frost, L. N., Richardson, H., Daniels, M. J., Jones, J. D. G., and Parker, J. E. Four *Arabidopsis RPP* loci controlling resistance to the *Noco2* isolate of *Peronospora parasitica* map to regions known to contain other *RPP* recognition specificities. *Molecular Plant-Microbe Interactions*, 9: 464–473, 1996.

51. Goss, E. M. and Bergelson, J. Variation in resistance and virulence in the interaction between *Arabidopsis thaliana* and a bacterial pathogen. *Evolution*, 60:1562–1573, 2006.

52. Alonso-Blanco, C., Bentsink, L., Hanhart, C. J., Vries, H. B. E., and Koornneef, M. Analysis of natural allelic variation at seed dormancy loci of *Arabidopsis thaliana*. *Genetics*, 164:711–729, 2003.

53. Cadman, C. S. C., Toorop, P. E., Hilhorst, H. W. M., and Finch-Savage, W. E. Gene expression profiles of *Arabidopsis* Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. *Plant Journal*, 46:805–822, 2006.

54. Melzer, S., Kampmann, G., Chandler, J., and Apel, K. *FPF1* modulates the competence to flowering in *Arabidopsis*. *Plant Journal*, 18:395–405, 1999.

55. Toh, S., Imamura, A., Watanabe, A., Nakabayashi, K., Okamoto, M., Jikumaru, Y., Hanada, A., Aso, Y., Ishiyama, K., Tamura, N., Iuchi, S., Kobayashi, M., Yamaguchi, S., Kamiya, Y., Nambara, E., and Kawakami, N. High temperature-induced abscisic acid biosynthesis and its role in the inhibition of gibberellin action in *Arabidopsis* seeds. *Plant Physiology*, 146: 1368–1385, 2008.

56. Hu, Y. X., Wang, Y. H., Liu, X. F., and Li, J. Y. *Arabidopsis RAV1* is down-regulated by brassinosteroid and may act as a negative regulator during plant development. *Cell Research*, 14:8–15, 2004.

57. Wu, G. and Poethig, R. S. Temporal regulation of shoot development in *Arabidopsis thaliana* by *miR156* and its target *SPL3*. *Development*, 133(18):3539–3547, 2006.

58. Gonzalez, D., Bowen, A. J., Carroll, T. S., and Conlan, R. S. The transcription corepressor *LEUNIG* interacts with the histone deacetylase *HDA19* and mediator components *MED14 (SWP)* and *CDK8 (HEN3)* to repress transcription. *Molecular And Cellular Biology*, 27:5306–5315, 2007.

59. Tominaga, R., Iwata, M., Sano, R., Inoue, K., Okada, K., and Wada, T. *Arabidopsis CAPRICE-LIKE MYB 3 (CPL3)* controls endoreduplication and flowering development in addition to trichome and root hair formation. *Development*, 135:1335–1345, 2008.

60. Laubinger, S., Marchal, V., Gentilhomme, J., Wenkel, S., Adrian, J., Jang, S., Kulajta, C., Braun, H., Coupland, G., and Hoecker, U. *Arabidopsis SPA* proteins regulate photoperiodic flowering and interact with the floral inducer *CONSTANS* to regulate its stability. *Development*, 133:4608–4608, 2006.

61. Araki, T., Kobayashi, Y., Kaya, H., and Iwabuchi, M. The flowering-time gene *FT* and regulation of flowering in *Arabidopsis*. *Journal Of Plant Research*, 111:277–281, 1998.

62. Doyle, M. R., Bizzell, C. M., Keller, M. R., Michaels, S. D., Song, J. D., Noh, Y. S., and Amasino, R. M. *HUA2* is required for the expression of floral repressors in *Arabidopsis thaliana*. *Plant Journal*, 41:376–385, 2005.

63. Western, T. L., Cheng, Y. L., Liu, J., and Chen, X. M. *HUA ENHANCER2*, a putative DExM-box RNA helicase, maintains homeotic B and C gene expression in *Arabidopsis*. *Development*, 129:1569–1581, 2002.

64. De Lucia, F., Crevillen, P., Jones, A. M. E., Greb, T., and Dean, C. A PHD-Polycomb Repressive Complex 2 triggers the epigenetic silencing of *FLC* during vernalization. *Proc. Natl. Acad. Sci. USA*, 105:16831–16836, 2008.

65. Kirik, V., Simon, M., Wester, K., Schiefelbein, J., and Hulskamp, M. *ENHANCER of TRY and CPC 2 (ETC2)* reveals redundancy in the region-specific control of trichome development of *Arabidopsis*. *Plant Molecular Biology*, 55:389–398, 2004.

ACKNOWLEDGMENTS

We thank B. Carvalho for his advice on how to modify the Oligo package. This work was primarily supported by NSF DEB-0519961 (J. Bergelson, M.N.), NIH GM073822 (J. Borevitz), and NSF DEB-0723935 (M.N.). Additional support was provided by the Dropkin foundation (J. Bergelson), NIH GM057994 and NSF MCB-0603515 (J. Bergelson), the Max Planck Society (D.W., M.T.), the Austrian Academy of Sciences (M.N.), the University of Lille 1 (F.R.), NIH GM62932 (J.C. and D.W.) and Howard Hughes Medical Institute (J.C.), DFG SFB 680 (J.d.M.), the Max Planck Society and a Gottfried Wilhelm Leibniz Award of the DFG (D.W.).

AUTHOR CONTRIBUTIONS

J.R.E. and D.W. generated the SNPs used by this project. S.A., M.H, Y.L., N.W.M, X.Z, J.Borevitz, and J.Bergelson were responsible for the experimental aspects of genotyping. Y.H., B.J.V., M.H., T.T.H., R.J., X.Z., M.A.A., P.M., J.Borevitz, J.Bergelson, 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.Bergelson, and M.N. were responsible for phenotyping. S.A., Y.H., B.J.V., G.W., D.M., A.P., A.T., P.M., and M.N carried out the GWA analyses. Y.H. and D.M. developed the project website. M.N. wrote the paper with major contributions from S.A., Y.H. B.J.V., G.W., A.P., and J.Bergelson. J.Borevitz, J.Bergelson, and M.N conceived and supervised the project.