A Nonparametric Test Reveals Selection for Rapid Flowering in the Arabidopsis Genome

Christopher Toomajian1*, Tina T. Hu1, Maria José Aranzana1u, Clare Lister2, Chunlao Tang1, Honggang Zheng1, Keyan Zhao1, Peter Calabrese1, Caroline Dean2, Magnus Nordborg1

Introduction

The detection of footprints of natural selection in genetic polymorphism data is fundamental to understanding the genetic basis of adaptation, and has important implications for human health. The standard approach has been to reject neutrality in favor of selection if the pattern of variation at a candidate locus was significantly different from the predictions of the standard neutral model. The problem is that the standard neutral model assumes more than just neutrality, and it is almost always possible to explain the data using an alternative neutral model with more complex demography. Today’s wealth of genomic polymorphism data, however, makes it possible to dispense with models altogether by simply comparing the pattern observed at a candidate locus to the genomic pattern, and rejecting neutrality if the pattern is extreme. Here, we utilize this approach on a truly genomic scale, comparing a candidate locus to thousands of alleles throughout the Arabidopsis thaliana genome. We demonstrate that selection has acted to increase the frequency of early-flowering alleles at the vernalization requirement locus FRIGIDA. Selection seems to have occurred during the last several thousand years, possibly in response to the spread of agriculture. We introduce a novel test statistic based on haplotype sharing that embraces the problem of population structure, and so should be widely applicable.

Citation: Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, et al. (2006) A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol 4(5): e137. DOI: 10.1371/journal.pbio.0040137

Academic Editor: Nick Barton, University of Edinburgh, United Kingdom

Received October 3, 2005; Accepted February 28, 2006; Published April 25, 2006

Copyright: © 2006 Toomajian et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: Col, Columbia; EHH, extended haplotype homozygosity; FRI, FRIGIDA; LD, linkage disequilibrium; Ler, Landsberg erecta; PHS, pairwise haplotype sharing

* To whom correspondence should be addressed. E-mail: cmtoomaj@alumni.uchicago.edu

¤ Current address: Department de Gene`tica Vegetal, Consorci Consejo Superior de Investigaciones Cientı´ficas-Institut de Recerca i Tecnologia Agroalimenta `ries, Barcelona, Spain

1 Molecular and Computational Biology, University of Southern California, Los Angeles, California, United States of America, 2 Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, Colney, Norwich, United Kingdom

The coordination of flowering time with environmental factors is a major determinant of reproductive success in plants, and numerous experiments in many different species have shown that populations are typically strongly adapted to their local environment. Plants decide when to flower by integrating multiple environmental and endogenous inputs: the underlying pathways are best understood in Arabidopsis thaliana, which has become the major model for studying flowering time [1]. A. thaliana occurs throughout the Northern Hemisphere in a wide range of environments, and shows considerable variation for flowering time [2–10]. In particular, many genotypes are extremely late-flowering unless exposed to prolonged cold temperatures—so-called vernalization—which likely functions to ensure a winter-annual habit. Several studies have shown that the locus FRIGIDA (FRI) explains much of this variation [5–10]. Recessive loss-of-function alleles at this locus essentially eliminate the requirement for vernalization [11]. Two such alleles, friCol and friLer (originally defined in the common laboratory accessions Columbia [Col] and Landsberg erecta [Ler]), account for a large proportion of early-flowering accessions found in Europe (Figure 1). Given their strong effect on a trait likely to be under strong selection, there is every reason to believe that these alleles have been under selection. Consistent with this, a study of polymorphism in the FRI chromosomal region suggested that they were both associated with long haplotypes (i.e., for each allele, most or all individuals carrying that allele also shared identical haplotypes over long chromosomal regions), possibly as a result of selection [12]. However, without a genomic control, there was no way of determining whether the length of these haplotypes was indeed unusual.

Here, we evaluate the evidence for recent selective sweeps on friCol and friLer by investigating patterns of haplotype sharing throughout the genome. Formally, we seek to reject the null hypothesis that the two loss-of-function alleles do not exhibit extremely high haplotype sharing (as would be expected if they had not been affected by recent selection), in favor of the alternative hypothesis that they do [13]. We use polymorphism data from 96 individual specimens of A. thaliana for which 1,102 short fragments have been sequenced as part of an ongoing survey of genomic polymorphism [14]. The density of our sequenced regions along each chromosome is sufficient to identify unusually long identical haplotypes among subsets of these individuals.

The rationale behind our approach is that chromosomes that are identical by descent at a polymorphic site must also share a short region surrounding that site. The length of this identical by descent region is influenced primarily by the age...
of the shared allele at the polymorphic site and the recombination rate in the region [15,16]. A mutation that is driven by directional selection will typically have reached its current population frequency much faster than a mutation that reached the same frequency as a result of genetic drift alone [17]. This leaves less time for recombination and mutation to break up the ancestral haplotype, resulting in a larger region of identity by descent around the mutation than expected given its current frequency [18–21]. Several studies have suggested methods for detecting this signature of recent selection [22–24]. A limitation common to all is that they rely on population genetics models to determine significance, models that are based on numerous assumptions about the demography and geographic structure of the population. In contrast, we take a nonparametric approach to detecting selection, and determine significance by comparing haplotype sharing around \(FRI\) with what is observed in the rest of the genome [25,26]. Our analysis shows that haplotype sharing around the two major \(FRI\) loss-of-function alleles is indeed unusually high, thus providing strong evidence for selection on these alleles.

**Results/Discussion**

In order to compare haplotype sharing around the loss-of-function alleles at \(FRI\) with sharing at thousands of reference loci throughout the genome, we developed a novel haplotype-sharing measure. In particular, we tried to develop a measure that accounts for population structure, because our sample is heavily structured [14]. Thus, pairs of individuals from the same population are more likely to share long haplotypes because they are more closely related than those from different populations. For a formal definition of our pairwise haplotype-sharing score (PHS), and details on how we calculate it, see Materials and Methods. Informally, the basis of this measure is the estimation of the shared length around any allele \(A\) at position \(x\) from a simple pairwise comparison between individual haplotypes [27]. Thus, haplotype sharing relative to allele \(A\) can only be measured if this allele is found on at least two haplotypes. Note that, due to inbreeding, comparisons can be made directly between individuals (i.e., haplotypes need not be inferred from genotypes). To account for population structure and increase our power to detect true footprints of selection, we normalized the length of identity of a pair of individuals around each allele by subtracting the mean length of identity found in these two individuals (across all polymorphic loci) and dividing by the standard deviation. Then, we averaged these normalized lengths over all pairs of individuals, regardless of which allele they carry. As a result, our measure misses a selective sweep that has affected the whole sample, and has reduced power to detect a sweep that has affected most of it: it is geared towards identifying alleles that are present in a minority of individuals.

**Haplotype Sharing around \(FRI\) Is Extreme**

We calculated PHS for all alleles in our data set where, at the corresponding polymorphic site, the minor allele is found in at least two accessions and we have data from at least 60 accessions. We also calculated the measure for \(fri_{Col}\) and \(fri_{Ler}\), which have been genotyped in our sample. The distribution of this measure, plotted against the frequency of each allele, is shown in Figure 2. Often, nearby alleles are highly correlated and provide redundant information about a long shared haplotype. Therefore, we eliminated all but the highest-

---

**Figure 1.** The Effect of Geography and \(FRI\) Alleles on Flowering Time
The map displays the position of 168 accessions from Europe, northern Africa, and western Asia. The histogram inset shows the distribution of the number of leaves at flowering without vernalization for these accessions. In both map and histogram, accessions carrying the \(fri_{Col}\) or \(fri_{Ler}\) alleles are indicated in red or blue, respectively, while the shade of gray in the remaining accessions corresponds to the number of leaves (i.e., 0–20). Bin sizes are halved (i.e., 0–10 leaves) in the histogram relative to the map.

DOI: 10.1371/journal.pbio.0040137.g001
loci would result in inflated phu values in a sizable number of reference statistical significance. Only a selective force that systematically decreases the genome-wide distribution of PHS to higher values, making it less likely that an individual candidate locus will reach genome-wide significance, will make the test conservative. Examples of recent selection among reference loci will shift the distribution of PHS in our reference loci, and we also do not believe that low PHS scores can be used here to identify ancient alleles preserved by long-term balancing selection. It is possible that a haplotype-sharing measure such as ours would perform better with higher-density polymorphism data, but we are skeptical that such an approach would be more powerful than those that use polymorphism-to- divergence ratios. The issue deserves further study.

EHH and the Effect of Population Structure

We also calculated haplotype sharing using a modification of the extended haplotype homozygosity (EHH) measure [23,28], which makes no correction for population structure. Briefly, for each allele where we have calculated PHS, we calculated EHH at flanking markers of increasing distance. We interpolated the distance at which EHH drops below 0.5 and used this as a score. We similarly removed all but the highest scoring alleles from nearby highly correlated alleles. The distribution of this score against allele frequency is superficially similar to that of PHS (not shown). Again, the friCol allele is above the 99th percentile given its frequency, and the friLer allele is just above the 95th percentile.

In general, however, alleles that exhibit extreme haplotype sharing using PHS do not necessarily do so using EHH (and vice versa). Because the PHS measure makes adjustments for some pairs of individuals being more related than others (see Materials and Methods), while EHH does not, we expect the latter measure to sometimes identify haplotype sharing that can be attributed to population structure rather than selection. To investigate this, we compared alleles with extreme values with respect to which individuals carried them. Separately for the PHS and EHH analyses, we clustered all alleles above the 99th percentile using unweighted pair group method with arithmetic averages (UPGMA) and a distance metric that measures whether alleles tend to be carried by the same individuals or not. The distance metric we used, the binary distance, treats each allele as a vector of binary bits (corresponding to the accessions), where the value of each bit (0 or 1) indicates whether the corresponding accession carries the given allele. The distance between two alleles is then the number of bits for which one allele is 0 and the other 1, divided by the total number of bits with at least one 1. To reduce noise, we restricted the analysis to alleles found in more than two individuals, and at frequency less than 0.5. If population structure were responsible for haplotype sharing, we would expect the same individuals to share haplotypes at several loci and the relevant alleles would be clustered. Indeed, in the EHH analysis, we found seven sets of low-frequency alleles that clustered with an average distance less than or equal to 0.5. Three pairs of alleles and one group of five alleles clustered with an average distance of less than 0.4 (Figure S1). In one extreme case, two alleles were
found in exactly the same set of accessions. The clustered alleles tend to be found in restricted geographic regions. For example, five alleles found exclusively in the United States cluster together, and four pairs and one triplet of alleles found predominantly in northern Sweden cluster together. This is consistent with the high haplotype sharing of these alleles resulting solely from the population structure present in the sample. In contrast, only three clusters of low-frequency alleles resulted from the PHS analysis, and just a single pair of alleles clustered at a distance of less than 0.4. These results demonstrate that, as expected, the EHH analysis is much more susceptible to identifying multiple alleles from the same sets of closely related individuals than the PHS measure when population structure is present in the sample. However, for organisms with a less extensive population structure than *A. thaliana*, such as humans, we expect the methods will yield very similar results.

The History of the FRI Alleles

What does the pattern of haplotype sharing and linkage disequilibrium (LD) around *friCol* and *friLer* tell us about the history of these alleles? Estimates can be obtained by counting the proportion of chromosomes carrying the allele that retain the inferred ancestral haplotype at linked markers. Following the approach of Stephens et al. [29], and treating fragments flanking FRI as multiallelic markers, we obtained rough age estimates of approximately 800 generations for *friCol* and 3,200 generations for *friLer*. The greater estimated age of *friLer* reflects the fact that haplotype sharing among these alleles is less extreme than for *friCol* and is consistent with the greater geographic spread of the former allele (Figure 1). More exact estimates are not possible without better estimates of the rates of outcrossing and recombination, and more information on population structure in *A. thaliana*, all of which affect the decay of LD. Our age estimates are inversely proportional to the rates of recombination in this region and historical outcrossing in this species. Therefore, if we assume we have overestimated either of these rates by a factor of two and recalculate age after halving the inflated rate, the ensuing age estimates would double. Nevertheless, assuming at least one generation per year for these rapid-cyclers, our estimates suggest that selection for weediness imposed by agriculture. Two other observations support the notion that the FRI-null alleles represent an adaptation to some form of human disturbance: first, flowering time variation in *A. thaliana* seems to be much more strongly correlated with geography (in particular latitude) if individuals carrying FRI-null alleles are excluded than if they are not excluded [8]; second, while *A. thaliana* accessions in general show strong isolation by distance, accessions carrying *friCol* found far from Germany (the center of distribution of this allele) generally are genetically similar to other German accessions [14], suggesting recent dispersal. More research is required to clarify the exact mechanism of selection.

**Scanning the Genome for Recent Selection**

Thus far we have demonstrated how the genomic distribution of haplotype sharing may be used in a priori tests of selection when strong candidates, such as the *FRI* loss-of-function alleles, exist. We now turn to a fundamentally different application, namely scanning the genome for signs of recent selection. Because we have no a priori candidates, we are simply trying to identify those alleles and regions that exhibit the strongest signals of selection [31,32]. However, since we are not assuming any statistical model, we have no way of assessing the significance of our findings. The top 1% is just that. Nonetheless, assuming that some recent selection has occurred in the genome, we expect that this set of alleles and their associated haplotypes will be enriched for regions affected by recent selection, and thus each of these regions can be considered as a candidate for recent selection.

In addition to *friCol*, 31 alleles found in more than two accessions and at frequency less than 0.8 were identified as clearly above the 99th percentile of PHS (Table S1). They were found on all five chromosomes, though they were not distributed evenly according to genetic length (figure 2 degrees of freedom, p < 0.005), with an excess found on Chromosomes 2 and 4. Some of these alleles, hereafter referred to as candidate alleles, may be associated with targets of selection; most are probably due to chance. We certainly plan to investigate a few of the more interesting regions further.

A problem when using haplotype sharing to scan the genome for selection is that it is difficult to identify the precise target of selection. Given that our polymorphism data represent approximately 0.5% of the genome, it is likely that the targets of selection reside in the 99.5% of the genome between our sequenced fragments and that most of the 31 candidate alleles merely are in strong LD with them. But even if we had complete sequence information, it would be difficult to identify the targets of selection. This is because selection at a single allele will often generate a complex pattern of increased haplotype sharing around multiple alleles throughout the chromosomal region. We attempted to remove alleles from our analyses that were obviously part of the same event. For instance, alleles in two fragments linked to *FRI* that were in strong LD with the *friCol* allele had PHS scores that are above the 99th percentile, but because these alleles were redundant with *friCol* and scored lower, they were removed. Yet our criteria for identifying redundant alleles are arbitrary, and will not always work.

Our analysis identified two regions with complex patterns of haplotype sharing, such as those detailed above, involving at least five candidate alleles within a single Mb. The most extreme example is Chromosome 2, where nine candidate alleles were found between positions 9.2 and 11.6 Mb. The other example is the short arm of Chromosome 4, where five candidate alleles were found between 0.2 and 1.2 Mb and a sixth was just short of the centromere at 2.4 Mb. The *FRI* gene is located at 0.2 Mb, and it is illuminating to consider the role of *FRI* in the overall pattern of haplotype sharing in this region. In Figure 3, we have plotted the location and extent of regions of haplotype sharing along Chromosome 4 (for the remaining chromosomes, see Figures S2–S5). On the short arm, four candidate alleles were minor alleles, and two of these, found at 1.1 Mb and 2.4 Mb, shared more than a single accession in common with *friCol*. The allele at 2.4 Mb is
associated with a haplotype that encompasses a reported pericentric inversion between the Columbia and Landsberg accessions [33]. This inversion, located at the opposite end of the chromosome arm from \( \text{FRI} \), is expected to interfere with recombination between two haplotype classes, and may result in errors in our estimates of both distance and haplotype sharing in the immediate region. Although the \( \text{friCol} \) allele was carried by two individuals that have no mismatches in our data through the inversion, the remaining individuals also displayed high levels of haplotype sharing, indicating that the haplotype sharing around \( \text{friCol} \) cannot be explained by the inversion. For \( \text{friCol} \) and the alleles at 1.1 Mb and 2.4 Mb, we removed the accessions that carried each allele in turn, and recalculated PHS for the other two alleles. In each case, the remaining two alleles were still extreme given their observed frequency, indicating that haplotype sharing around each allele cannot be explained by the other alleles. The remaining minor frequency alleles appeared to be independent of \( \text{friCol} \) but it is impossible to rule out selection on this allele having induced extensive haplotype sharing among other accessions. The final candidate allele in this region, a major frequency allele, was too common to test its independence from other alleles in this region by removing all accessions that carry it. A detailed view of haplotype structure along the short arm of Chromosome 4 is provided in Figure S6.

**Concluding Remarks**

Taken together, the effect, frequency, and geographic distribution of the two most common \( \text{FRI} \) loss-of-function alleles had strongly suggested that they had been involved in adaptation. Our analysis provides strong support for this conjecture for at least one of these alleles, \( \text{friCol} \). The evidence for selection on the other allele, \( \text{friLer} \), is weaker: given that the two alleles have very similar phenotypic effects, the simplest explanation for this is that the selective sweep took place longer ago. Our genome-wide scan for selection has also identified many other regions that are candidates for harboring genes that have been the subject of recent selective sweeps. As we learn more about the genes responsible for natural variation in this organism, it will be interesting to see how many lie in these putatively selected chromosomal regions. The type of scan developed here should be useful in a wide range of species with cryptic population structure, and with some modification for genotype data, could be applied to data collected by the Human International Haplotype Mapping Project.

**Materials and Methods**

**Genome-wide polymorphism dataset and sample choice.** The majority of data used here has already been described [14], and the rest were gathered in the same manner. Briefly, we have sequenced 1,102 fragments of approximately 500 bp in a set of 96 accessions. The mean and median distance between adjacent fragments are 107.5 and 96.6 kb, respectively, or an estimated 0.49 and 0.40 cM, respectively. Full alignments for these fragments are provided in Dataset S1. The 96 accessions were chosen for the previous study both to capture some information on within-region variation in \( \text{Arabidopsis} \) as well as include accessions that have been commonly used in the creation of recombinant inbred lines. The fact that we account for population structure in our approach implies that our results should be robust to sample composition. We took the sampled frequency of each allele as an estimate of its global frequency. As the PHS score

**Figure 3. Haplotype Sharing of Alleles Plotted by Frequency along Chromosome 4**

Points represent the location of alleles, while corresponding lines indicate the average length of haplotype identity flanking the allele. Those alleles present in the upper 99th percentile are indicated in black. The \( \text{friCol} \) allele and associated haplotype are indicated in red. This haplotype is over 600 kb, or 8 cM, long. The allele marking \( \text{friLer} \) and its associated haplotype are indicated in blue. Only those alleles with a positive score are plotted. Locations of the sequence fragments are indicated at the bottom of the figure.

DOI: 10.1371/journal.pbio.0040137.g003
shows a decline with increasing frequency, errors in these global frequency estimates could affect our results. However, accessions were not chosen on the basis of FRI genotypes.

**PHS.** We defined the haplotype-sharing statistic

$$PHS_x = \frac{\sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{(i \neq j)} Z_{ij} \sum_{(i \neq j)} Z_{ij}}{\binom{p}{2} \binom{p}{2}}$$

where $p$ is the number of individuals carrying allele $A$ at position $x$, and $n$ is the sample size at position $x$; and

$$Z_{ij} = \frac{d_{ijx} - \bar{d}_{ijx}}{\sigma_{ij}}$$

where $d_{ijx}$ is the distance on a genetic map over which individuals $i$ and $j$ are identical around position $x$, $\bar{d}_{ijx}$ is the mean of the genome-wide distribution of distances over which these individuals are identical, $\sigma_{ij}$ is the standard deviation of that distribution, and individuals $i$ and $j$ share the same allele at position $x$. The normalized genetic distance $Z_{ij}$ was employed in order to control for population structure in our data. The final statistic is expressed as the difference of two averages (the first, averaged over all pairs that share the allele; the second, over all possible pairs) to guard against genome regions where in average average pairs of accessions are identical for long stretches, either from few sequenced fragments, or few or low-frequency polymorphisms in the available sequences.

Heterozygous sites (rare in our data) were converted to missing data. Missing data values do not interrupt haplotype identity between pairs of individuals, nor do they count toward extending identical haplotypes until they are followed by a polymorphism where both individuals are identical.

For each allele we also calculated the average genetic distance to the left and right that any two individuals carrying the allele are identical and recorded these as the endpoints of sharing for that allele. Because our polymorphism data represents only a fraction of the whole genome, we can usually only infer the location where identical haplotypes first differ. To be conservative, we measured haplotype identity out to the last whole adjacent fragment that is identical between the pair of individuals. These endpoints of haplotype sharing are approximate and serve only as a guide to the extent of individual regions of haplotype sharing.

**Filtering highly correlated alleles.** LD between nearby alleles is often very high. Because we exhaustively calculate PHS for each allele in our dataset for all polymorphisms that have genotypes from at least 60 individuals, this often leads to the same score for many nearby alleles that are perfectly correlated. In this case, each of these alleles represents the same block of haplotype sharing. To reduce over-counting blocks of haplotype sharing, we removed all but the highest scoring allele from each group of correlated alleles. Correlated alleles are those between which $r^2 > 0.5$ and $r$ is positive, and they are within the same fragment or their intervals of haplotype sharing overlap. We expect this filtering reduces the amount of redundancy in the high scoring alleles, though it is likely that pairs or groups of alleles still remain that are extreme due to the same shared haplotypes.

**EEH analysis.** As EHH for a given allele or haplotype can be calculated at many flanking markers, it is not obvious how to compare this measure with PHS, which has a unique value. To make PHS and EHH more comparable, we calculated a score based on EHH as follows. For each allele where we have calculated PHS, we also calculated EHH (and EHH for all individuals who do not carry the given allele) at flanking polymorphic sites increasingly distant from the allele. Independently to the left and right, we found the polymorphic sites between which EHH drops below 0.5 and interpolated the position where this occurs. The score is then the total distance on a genetic map between the two positions thus estimated [28]. The score was assigned to zero for all alleles where this genetic distance is less than or equal to the genetic distance between the sites where EHH measured for all individuals who do not carry the allele drops below 0.5.

**Estimates of distances on a genetic map.** Functions to translate physical distances into distances on a genetic map (i.e., recombination frequencies) were created for each chromosome. Briefly, a third-order polynomial was fitted to data on position on physical and genetic maps for markers on each chromosome arm. We used a total of 483 markers from http://www.arabidopsis.org, where positions on both physical and genetic maps are available. In Supporting Information we include Table S2 describing these markers, as well as Figures S7–S11, which illustrate the fitted curves of genetic by physical distance for each chromosome and the markers used to produce those curves. For the first arm of each chromosome, the polynomial was forced through the origin. In the physical region between chromosome arms representing each centromere, an estimate of the genetic position of the centromere was applied. The regions that approximated the physical extent of each centromere were given a zero genetic distance. In order that the function may be monotonically increasing, a short linear segment was added between the first arm of Chromosome 4 and its centromere, where the fitted polynomial started to decrease.

**Age estimates of alleles.** Following Equation 2 of Stephens et al. [29], we obtained estimates of the time to the most recent common ancestor of the friLer and friCol alleles from the breakdown of LD at flanking fragments. We assigned the site as the ancestral haplotype of each FRI allele at each fragment as the majority haplotype among the individuals that were identical from the core allele to the previous fragment. In estimating $r$, the rate of change from the ancestral haplotype at each fragment, we used the genetic distance from FRI (a proxy for recombination fraction) multiplied by the proportion of recombination events resulting from outcrossing that would result in the loss of the ancestral haplotype (i.e., the proportion of chromosomes not carrying the specified deletion that do not carry the deletion’s ancestral haplotype) and the rate of outcrossing, here taken to be 0.01. Uncertainty exists in the rates of both recombination and outcrossing, and this uncertainty is difficult to quantify [34], though present day outcrossing estimates from disturbed populations may greatly overestimate historical outcrossing rates. These rates are linearly proportional to $r$, which is inversely proportional to the estimated time, meaning our estimates of the time to the most recent common ancestor from estimates at seven and nine flanking fragments, respectively.

**Supporting Information**

**Dataset S1.** All Data Used in This Paper, Including Genome-Wide Sequence Alignments, FRI Genotypes, and Flowering Time Data

Found at DOI: 10.1371/journal.pbio.0040137.sd001 (1.79 MB ZIP).

**Figure S1.** Clustering of High-Scoring Alleles from PHS and EHH Analyses

Clustering is based on the similarity among alleles of which individuals carry them (measured by binary distance). Clustering of alleles from the EHH analysis is presented at the top of the figure, while PHS clustering is below. Colored boxes highlight clusters where the mean binary distance between alleles is less than or equal to 0.5. To the right of the clusters are color-coded labels describing the regions where the boxed alleles predominate.

Found at DOI: 10.1371/journal.pbio.0040137.sg001 (28 KB PDF).

**Figure S2.** Haplotype Sharing of Alleles Plotted by Frequency along Chromosome 1

Found at DOI: 10.1371/journal.pbio.0040137.sg002 (11.6 MB PDF).

**Figure S3.** Haplotype Sharing of Alleles Plotted by Frequency along Chromosome 2

Found at DOI: 10.1371/journal.pbio.0040137.sg003 (146 KB PDF).

**Figure S4.** Haplotype Sharing of Alleles Plotted by Frequency along Chromosome 3

Found at DOI: 10.1371/journal.pbio.0040137.sg004 (5.6 MB PDF).

**Figure S5.** Haplotype Sharing of Alleles Plotted by Frequency along Chromosome 5

Found at DOI: 10.1371/journal.pbio.0040137.sg005 (6.4 MB PDF).

**Figure S6.** Haplotype Structure of the Short Arm of Chromosome 4

For each alignment in this region, nonredundant (i.e., $r^2 = 1$) polymorphisms are indicated in columns and accessions in rows. Minor frequency alleles are black and major frequency alleles are gray, while missing data or spaces between alignments are white. A genetic scale (in cm) is included at the top of the figure, and the position of each alignment on it is indicated. The extended haplotypes surrounding seven alleles that are in the upper 2.5% tally of genome-wide PHS and that are below 15% frequency are boxed, and the fragment where the core allele is located is also boxed. The haplotype surrounding friLer is boxed in red, and that surrounding the
marker allele for fril1 in blue. Due to topological constraints, one accession belonging to the third group from the top is not shown in this group because it is included with the fril1 marker group.

Found at DOI: 10.1371/journal.pbio.0040137.sg006 (5.1 MB JPG).

**Figure S7.** Fitted Curves Relating Genetic and Physical Positions for Chromosome 1
Lines represent third-order polynomials fitted to markers (points) for each chromosome arm and the centromere. The density of sequence alignments is indicated along both axes. Found at DOI: 10.1371/journal.pbio.0040137.sg007 (73 KB PDF).

**Figure S8.** Fitted Curves Relating Genetic and Physical Positions for Chromosome 2
Found at DOI: 10.1371/journal.pbio.0040137.sg008 (56 KB PDF).

**Figure S9.** Fitted Curves Relating Genetic and Physical Positions for Chromosome 3
Found at DOI: 10.1371/journal.pbio.0040137.sg009 (61 KB PDF).

**Figure S10.** Fitted Curves Relating Genetic and Physical Positions for Chromosome 4
Found at DOI: 10.1371/journal.pbio.0040137.sg010 (58 KB PDF).

**Figure S11.** Fitted Curves Relating Genetic and Physical Positions for Chromosome 5
Found at DOI: 10.1371/journal.pbio.0040137.sg011 (63 KB PDF).

**Table S1.** Description of Alleles Associated with High Haplotype Sharing
Found at DOI: 10.1371/journal.pbio.0040137.st001 (22 KB XLS).

**Table S2.** List of Genetic Markers Used
Found at DOI: 10.1371/journal.pbio.0040137.st002 (91 KB XLS).

**Accession Numbers**

The Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) accession number for FRI is 828044.

**Acknowledgments**

We thank Liz Cooper for assistance with data generation, and Norm Arshady, Jeff Wall, and the anonymous reviewers for constructive comments on the manuscript.

**Author contributions.** CT, TTH, PC, and MN conceived and designed the experiments. CT, TTH, KZ, and PC carried out the analysis. MJA, CL, CT, and HZ provided FRI genotype and flowering time data. CT and MN wrote the paper.

**Funding.** This work was mainly supported by the US National Science Foundation (NSF 2010 grant DBE-0115062 to MN, Martin Kreitman, and Joy Bergelson) and the UK Natural Environment Research Council (NER/T/S/2001/001240 to CD). In addition, HZ was supported by a grant from the Fletcher Jones Foundation (Simon Tavaré, PI), MJA was supported by a grant from the W. H. Keck Foundation (MN, PI), CT was supported by a postdoctoral fellowship from the National Institutes of Health, and PC was supported by Center of Excellence in Genomic Science grant P50 HG002790 (M. Waterman, PI) from the National Institutes of Health.

**Competing interests.** The authors have declared that no competing interests exist.

---

**References**

1. Simpson GG, Dean C (2002) Arabidopsis, the rosetta stone of flower time? Science 296: 285–289.
2. Karlsson BH, Sills GR, Nienhuis J (1993) Effects of photoperiod and vernalization on the number of leaves at flowering in 32 Arabidopsis thaliana (Brassicaceae) ecotypes. Am J Bot 80: 646–648.
3. Sauta SL, John M, Amasino RM (1997) Analysis of flowering time in Arabidopsis thaliana. Genetics 147: 995–1008.
4. Nordborg M, Bergelson J (1999) The effect of seed and rosette cold treatment on germination and flowering time in some Arabidopsis thaliana (Brassicaceae) ecotypes. Am J Bot 86: 470–475.
5. Gazzani S, Gundall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. Plant Physiol 132: 1107–1114.
6. Sanda SL, Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, et al. (2004) Haplotype structure and phenotypic associations in the chromosomal regions comprising two Arabidopsis thaliana flowering time loci. Genes 168: 1627–1638.
7. Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, et al. (2004) A pre-expansion CAG repeat of the human Fragile X (FMR1) locus. Genome Res 11: 1382–1391.
8. Yu F, Saberi PC, Hardenbol P, Fu Q, Fry B, et al. (2005) Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. PLoS Genet 1: e11. DOI: 10.1371/journal.pgen.0010041
9. Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A (2001) Sequence variation within the Fragile X locus. Genome Res 11: 1382–1391.
10. Saberi PC, Walsh E, Schaffner SF, Varilly P, Fry B, et al. (2005) The case for detection of recent selection in the human genome from haplotype structure. Nature 419: 832–837.
11. Toomajian C, Ajoka RS, Jurde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165: 287–297.
12. Krommern J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. Nature 435: 95–98.
13. Yu F, Saberi PC, Hardenbol P, Fu Q, Fry B, et al. (2005) Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. PLoS Genet 1: e11. DOI: 10.1371/journal.pgen.0010041
14. Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A (2001) Sequence variation within the Fragile X locus. Genome Res 11: 1382–1391.
15. Saberi PC, Walsh E, Schaffner SF, Varilly P, Fry B, et al. (2005) The case for detection of recent selection in the human genome from haplotype structure. Nature 419: 832–837.
16. Toomajian C, Ajoka RS, Jurde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165: 287–297.
17. Maruyama T (1974) The age of an allele in a finite population. Genet Res 23: 137–143.
18. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 25–35.
19. Kaplan NL, Hudson RR, Langley CH (1989) The ‘‘hitchhiking effect’’ revisited. Genetics 123: 887–899.
20. Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide-dismutase (SOD) region of Drosophila melanogaster. Genetics 136: 1329–1340.
21. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.
22. Slatkin M, Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics 158: 865–874.
23. Saberi PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
24. Toomajian C, Ajoka RS, Jurde LB, Kushner JP, Kreitman M (2003) A method for detecting recent selection in the human genome from allele age estimates. Genetics 165: 287–297.
25. Kroymann J, Mitchell-Olds T (2005) Epistasis and balanced polymorphism influencing complex trait variation. Nature 435: 95–98.
26. Yu F, Saberi PC, Hardenbol P, Fu Q, Fry B, et al. (2005) Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. PLoS Genet 1: e11. DOI: 10.1371/journal.pgen.0010041
27. Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A (2001) Sequence variation within the Fragile X locus. Genome Res 11: 1382–1391.
28. Saberi PC, Walsh E, Schaffner SF, Varilly P, Fry B, et al. (2005) The case for detection of recent selection in the human genome from haplotype structure. Nature 419: 832–837.
29. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, et al. (1998) Dating the origin of the CCR5-A32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62: 1507–1515.
30. Hewitt GM (2001) Speciation, hybrid zones and phylogeography—or seeing genes in space and time. Mol Ecol 10: 537–540.
31. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805–1814.
32. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
33. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, et al. (2000) Integrated cytogenetic map of Chromosome arm 4S in A. thaliana: Structural organization of euchromatinic knob and centromere region. Cell 100: 367–376.
34. Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. Heredity 62: 411–418.