J Mol Evol (2008) 66:116–129
DOI 10.1007/s00239-007-9063-3

Local Patterns of Nucleotide Polymorphism Are Highly Variable in the Selfing Species *Arabidopsis thaliana*

Richard C. Moore · M. Henry H. Stevens

Received: 5 March 2007 / Accepted: 21 November 2007 / Published online: 14 February 2008

Abstract Neighboring genes predictably share similar evolutionary histories to an extent delineated by recombination. This correlation should extend across multiple linked genes in a selfing species such as *Arabidopsis thaliana* due to its low effective recombination rate. To test this prediction, we performed a molecular population genetics analysis of nucleotide polymorphism and divergence in chromosomal regions surrounding four low-diversity loci. Three of these loci, At1g67140, At3g03700, and *TERMINAL FLOWER1* (*TFL1*), have been previously implicated as targets of selection and we would predict stronger correlations in polymorphism between neighboring loci due to genetic hitchhiking around these loci. The remaining locus, At1g04300, was identified in a study of linkage disequilibrium surrounding the *CRYPTOCHROME2* (*CRY2*) locus. Although we found broad valleys of reduced nucleotide variation around two of our focal genes, At1g67140 and At3g03700, all chromosomal regions exhibited extreme variation in the patterns of polymorphism and evolution between neighboring loci. Although three of our four regions contained potential targets of selection, application of the composite-likelihood-ratio test of selection in conjunction with a goodness-of-fit test supports the selection hypothesis only for the region containing At3g03700. The degree of discordance in evolutionary histories between linked loci within each region generally correlated with estimates of recombination and linkage disequilibrium for that region, with the exception of the region containing At1g04300. We discuss the implications of these data for future population genetics analyses and genomics studies in *A. thaliana*.

Keywords Plant · Population genetics · Nucleotide diversity · *TFL1* · Genomics

Abbreviations
TFL1 *(TERMINAL FLOWER1)*

Introduction

Understanding the evolutionary forces that shape patterns of nucleotide polymorphism remains a critical goal of molecular population geneticists. Genomic analyses of the patterns of polymorphism and divergence suggest that these values are generally conserved between linked loci; that is, neighboring loci have more similar levels of polymorphism and divergence than do random pairs of loci (Hahn 2006; Matassi et al. 1999). For example, in a survey of the Drosophila X chromosome, Hahn (2006) determined that this correlation extended up to 300 kb for polymorphism values and almost 100 kb for divergence values. This correlation also extends to the rate of protein evolution. In the mouse genome, the rate of protein evolution, as gauged by the ratio of nonsynonymous divergence (K_a) to synonymous divergence (K_s), is highly correlated between neighboring genes up to 1 cM (2 Mbp) apart (Williams and Hurst 2000).

Knowing the extent to which the patterns of polymorphism and divergence are related between linked loci has

Electronic supplementary material The online version of this article (doi:10.1007/s00239-007-9063-3) contains supplementary material, which is available to authorized users.

R. C. Moore (✉) · M. H. H. Stevens
Department of Botany, Miami University, 316 Pearson Hall,
Oxford, OH 45056, USA
e-mail: moorerc@muohio.edu

significant bearing on the ability to identify loci that potentially contribute to evolutionary change (e.g. Akey et al. 2002; Diller et al. 2002; Mousset et al. 2003; Nurminsky et al. 1998; Payseur et al. 2002; Shimizu et al. 2004; Tiffin et al. 2004; Vigouroux et al. 2002; Wiener et al. 2003). For example, positive selection acting on a newly-arisen beneficial mutation can cause dramatic decreases in nucleotide variation through a selective sweep, as the beneficial allele rises to fixation faster than mutation can introduce new alleles to the population. Alternatively, balancing selection can act to maintain divergent alleles, increasing nucleotide diversity at selected loci (Charlesworth 2006). Through the action of genetic hitchhiking, these signatures of selection predictably extend to neighboring loci to an extent dependent on the strength of selection and the recombination rate (Barton 2000; Kaplan et al. 1989; Maynard Smith and Haigh 1974). Demographic forces, such as population bottlenecks, can also affect nucleotide diversity levels; however, demographic forces affect genome-wide patterns of nucleotide variation. In contrast to both selection and demography, a stochastic decrease in molecular variation caused by genetic drift is more localized and does not necessarily extend to neighboring loci. Thus, when trying to infer the evolutionary history of a particular locus, it is important to quantify variation at linked loci.

The extent to which linked neutral variation is affected by selection is determined, in part, by recombination, which severs the association of linked loci and constrains the extent to which evolutionary histories between linked loci are shared. In a selfing species, such as the model plant *Arabidopsis thaliana*, the low effective recombination rate should increase the correlation of evolutionary histories between linked loci (Glemin et al. 2006). For example, the signature of selection surrounding the pseudo-*SCR1* locus involved in the loss of self-incompatibility spans up to 35 kb (Shimizu et al. 2004). However, even regions containing loci suspected of being targets of selection can exhibit variable patterns of polymorphism between neighboring loci (Hagenblad and Nordborg 2002; Haubold et al. 2002; Shepard and Purugganan 2003). In such cases gene conversion, versus reciprocal recombination, may play a more prominent role in structuring patterns of variation between neighboring loci (Haubold et al. 2002).

In order to assess the degree to which the evolutionary histories of linked loci are correlated in the *A. thaliana* genome, we examined the molecular population genetics of four genomic regions containing previously identified low-diversity genes. Although previous population genetics studies suggest that the low nucleotide diversity at three of our four focal genes is consistent with the action of positive selection (Barrier et al. 2003; Moore and Purugganan 2003; Olsen et al. 2002), there was considerable variation in the levels and patterns of nucleotide polymorphism in neighboring loci in all four genomic regions. Indeed, application of a test of selection which takes into account patterns of polymorphism and divergence of linked loci supported the selection hypothesis only for the region containing At3g03700. For three of the genomic regions, patterns of recombination and the extent of linkage disequilibrium correspond to the degree of heterogeneity in evolutionary histories between linked loci. However, one region that exhibits considerable dissimilarity between neighboring loci is found in a chromosomal region with the lowest estimates of recombination that exhibits extensive linkage disequilibrium. We discuss the implications of these data for future population genetics analyses and genomics studies in *A. thaliana*.

Materials and Methods

Isolation and Sequencing of Alleles

Genomic DNA was isolated from young leaves of 14 *A. thaliana* accessions (Supplementary Table 1) and one *Arabidopsis lyrata* individual using the Plant DNeasy Mini Kit (Qiagen, Valencia, CA). DNA was isolated from an additional 30 ecotypes for a subset of genes surrounding At1g04300 (Supplementary Table 1). The *A. lyrata* individual was grown from seed isolated from a Karhumaki, Russia population and was provided by O. Savolainen and Helmi Kuittinen (University of Oulu, Oulu, Finland). Between 0.7 and 1 kb of coding region was sequenced for each locus. In addition, previously published sequence data from additional ecotypes were incorporated when available (Barrier et al. 2003; Moore and Purugganan 2003; Olsen et al. 2004; Olsen et al. 2002).

PCR primers were designed based on the Col-0 gene sequences using Primer3 (Rozen and Skaletsky 2000); (Supplementary Table 2). PCR of *A. thaliana* and *A. lyrata* samples was performed with Taq DNA polymerase (Roche, Indianapolis, IN). DNA fragments amplified from *A. thaliana* were purified using the QIAquick Gel Extraction Kit (Qiagen) and directly sequenced. Amplified *A. lyrata* products were subcloned using the TA TOPO PCR Cloning Kit (Invitrogen, Carlsbad, CA), and plasmid DNA from six independent clones was sequenced. DNA sequencing was conducted at the NCSU Genome Research Laboratory with a Prism 3700 96-capilary automated sequencer (Applied Biosystems, Foster City, CA). All polymorphisms were visually confirmed and ambiguous polymorphisms were rechecked by PCR reamplification and sequencing. GenBank accession numbers for these genes are EU351021-EU352144.

Molecular Evolution and Population Genetic Data Analysis

Sequences were aligned against the *A. thaliana* sequence previously identified in the Arabidopsis whole genome sequence (Arabidopsis Genome Initiative 2000). The *A. lyrata* ortholog was used as the outgroup in the analyses. Levels of silent site (synonymous and noncoding) nucleotide diversity were estimated as π (Nei 1987) and θ_W (Watterson 1975). Nucleotide divergence at silent sites (K_{JC}) was determined with Jukes and Cantor correction (Jukes and Cantor 1969) according to Nei (1987). All estimates of polymorphism and divergence were determined using DnaSP 4.0 (Rozas and Rozas 1999). DnaSP 4.0 was also used to calculate the Tajima's D and Fay and Wu's H test statistic (Fay and Wu 2000; Tajima 1989) .

Statistical Methods

Testing for Correlation of Population Genetic Statistics Between Neighboring Loci

A randomization method based on the one Williams and Hurst (2000) employed to test the correlation of protein evolution rates between adjacent loci was used in order to assess the degree of similarity in population genetic statistics (π_{sil}, θ_W, θ/K, Tajima's D, and Fay and Wu's H) between neighboring loci within each region. This stands in contrast to studies in other organisms that have used autocorrelation analysis to test if neighboring loci have more similar polymorphism and divergence values (e.g. Hahn 2006, Matassi et al. 1999). For our dataset, neither autocorrelation analysis (Diggle 1990) nor spectral analysis (Bloomfield 2000) was appropriate, as such analyses have low power to detect patterns with small numbers of observations (SAS Institute 1990). Using our method, the difference in each statistic between pairs of neighboring loci in the four regions (e.g. $\Delta\pi_{sil}$, $\Delta\theta_W$,$\Delta[\theta/K]$, ΔD, ΔH) was first calculated. The means of these values for each region were then calculated, giving us a measure of the similarity in each statistic between neighboring loci. In order to assess the significance of these values, 10,000 random rearrangements of each statistic within a given region were produced. For each randomized data set, all values were randomly reallocated to positions within each region and the mean change in each statistic between neighboring loci was calculated (e.g. mean $\Delta\pi_{sil}$, mean $\Delta\theta_{W,}$ mean $\Delta[\theta/K]$, mean ΔD, mean ΔH). If there is a significant correlation in a statistic between neighboring loci, less than 5% of the random arrangements will have a lower mean difference than the observed value.

Testing for Selection in Each Region

The composite likelihood-ratio (CLR) test of Kim and Stephan (2002) as implemented in the program clsw was used to determine if the patterns of nucleotide variation in each region is consistent with a selective sweep. This program is available as part of the composite likelihood analysis (CLA) software package available at http://yuseobkim.net/YuseobPrograms.html. CLR analysis contains two options; option 1 (test A) distinguishes between derived and ancestral alleles and option 2 (test B) does not make this distinction. DnaSP 4.0 was used to estimate the recombination parameter (R) between adjacent sites (Hudson 1987). CLR values for each region were compared to those from 1000 replicate neutral data sets simulated using the program ssw (part of the CLA software package). Regions with significant CLR for either option ($P < 0.05$) were further analyzed by a goodness-of-fit (GOF) test described by Jensen et al. (2005), as implemented in the clsw program. The GOF test is designed to distinguish between selection and demographic forces which can give rise to similar patterns of variation. GOF values are compared to 1000 replicate data sets generated by selective sweep simulations using the program ssw. Regions with low GOF values ($P > 0.05$) are consistent with the selective sweep hypothesis.

Analysis of Recombination and Linkage Disequilibrium

Global estimates of recombination for regions containing focal low-diversity loci were estimated from the derivative of a fifth order polynomial of genetic versus physical distance for collinear markers from the Lister and Dean Col × L*er* recombinant inbred map (Lister and Dean 1993). Physical distances were obtained from ftp://ftp.arabidopsis.org/home/tair/Maps/mapviewer_data/. DNAsp 4.0 was used to estimate linkage disequilibrium (LD, as measured by the square of the correlation coefficient, r^2) between informative polymorphic loci within each genomic region. In order to assess LD between all loci within each region, we combined sequence alignments for the region into one sequence file. The intermittent regions between sequenced loci were treated as missing data. The minimum number of recombination events (R_m) was obtained using the four-gamete test (Hudson and Kaplan 1985). All recombination analyses were performed using DnaSP 4.0.

Results

Patterns of Genetic Diversity Can Vary Greatly Between Neighboring Loci

In a selfing species such as *A. thaliana*, with its reduced effective recombination, neighboring loci should have correlated evolutionary histories (Glemin et al. 2006). This effect will be exacerbated for loci surrounding the target of recent selection due to genetic hitchhiking (Barton 2000; Kaplan et al. 1989; Maynard Smith and Haigh 1974). In order to determine the extent to which neighboring loci share similar evolutionary histories in *A. thaliana*, we surveyed patterns and levels of nucleotide diversity in 1 kb portions of genes located at 5 to 10 kb intervals from four focal low-diversity loci and spanning chromosomal regions ranging from 43 to 113 kb (Table 1; Supplementary Table 3).

Table 1 Nucleotide diversity and divergence of loci in genomic regions containing focal genes

Gene	Position (kb)	n	Total sites (bp)	Silent sites (bp)	S^{a}	S^{b}_{sil}	θ^{c}_{W}	π^{d}_{sil}	K^{e}_{JC}	Tajima's D	Fay and Wu's H
A) Region I											
At1g67050	0	14	755	166	11	7	0.013	0.006	0.12	−2.15	−4.22
At1g67070	14.7	14	788	285	5	1	0.001	0.0005	0.24	−1.12	0.791
At1g67100	24.9	14	794	307	0	0	0	0	0.16	n.a.	n.a.
At1g67110	34.9	14	1001	462	33	32	0.022	0.033	0.19	2.24	−0.879
At1g67120a	42.1	14	1000	200	3	1	0.002	0.0007	0.11	−1.67	0.264
At1g67120b	52.3	14	1078	679	13	12	0.006	0.004	0.10	−1.33	0.484
At1g67120c	61.5	14	1148	354	4	2	0.002	0.0008	0.07	−1.80	0.527
At1g67140a[f]	72.5	23	866	431	5	1	0.0006	0.0002	0.10	−1.81	0.502
At1g67140b	81.9	14	1120	843	5	5	0.002	0.001	0.09	−0.873	−0.264
At1g67140c	87.5	14	969	635	6	6	0.003	0.002	0.12	−1.08	0.571
At1g67150	94.6	14	1096	299	32	27	0.028	0.033	0.13	0.754	0.044
At1g67170	99.2	14	1047	501	7	5	0.003	0.004	0.17	0.418	0.527
At1g67210	112.9	14	761	555	14	11	0.006	0.004	0.13	−1.77	−1.63
B) Region II											
At3g03660	0	14	857	439	8	4	0.003	0.003	0.09	0.067	0.835
At3g03670	10.7	14	773	264	2	2	0.002	0.001	0.12	−1.48	0.264
At3g03680	16.4	14	1063	244	17	10	0.013	0.011	n.d.	−0.623	n.d.
At3g03690	20.2	14	866	449	4	1	0.0007	0.0003	0.11	−0.243	0.791
At3g03700[f]	25	23	1026	643	0	0	0	0	0.08	n.a.	n.a.
At3g03720	35.3	14	1063	753	3	2	0.0008	0.0005	0.08	−0.886	0.571
At3g03730	44.2	14	797	179	7	1	0.002	0.002	0.05	0.142	0.264
At3g03740	45.9	14	1011	379	4	3	0.003	0.003	0.08	−0.186	0.528
At3g03770	54.5	14	795	503	4	4	0.003	0.002	0.11	−1.22	−1.32
At3g03780	66.5	14	806	419	3	1	0.0008	0.0006	0.09	−0.886	−2.51
At3g03790	75.6	14	766	349	15	8	0.007	0.008	0.16	0.881	−4.09
C) Region III											
At1g04190	0	14	638	483	5	5	0.003	0.004	0.11	0.797	−0.044
At1g04210	10.3	14	1057	493	4	4	0.003	0.003	0.15	−0.071	−0.308
At1g04220	13.9	14	722	177	5	4	0.007	0.006	0.18	−0.558	−1.98
At1g04240	20.7	44	777	335	6	3	0.002	0.0005	0.08	−1.55	0.522
At1g04250	28.8	44	876	680	7	7	0.002	0.002	0.15	−0.751	−0.416
At1g04270	34.1	44	1004	674	46	46	0.016	0.006	0.15	−2.26	−21.4
At1g04280	36.9	44	937	537	10	7	0.003	0.005	0.22	1.38	−1.79
At1g04290	39.9	44	562	184	10	9	0.008	0.014	0.13	1.76	0.660
At1g04300[f]	43.5	55	985	714	7	7	0.002	0.0005	0.08	−1.99	−1.69
At1g04310	48.3	44	999	233	6	2	0.002	0.002	0.12	−0.829	0.256
At1g04330	53.5	44	509	290	16	14	0.011	0.009	0.19	−0.745	−1.56

Table 1 continued

Gene	Position (kb)	n	Total sites (bp)	Silent sites (bp)	S[a]	S_{sil}[b]	θ_W[c]	π_{sil}[d]	K_{JC}[e]	Tajima's D	Fay and Wu's H
At1g04350	57.5	44	804	233	4	2	0.002	0.002	0.13	−0.486	−0.636
At1g04360	59.7	44	1049	234	12	2	0.002	0.0004	0.19	−0.681	1.44
At1g04370	67.4	44	397	93	3	1	0.003	0.0005	0.16	−0.821	0.336
At1g04380	69.5	32	906	348	22	8	0.006	0.006	0.14	0.117	−10.9
At1g04390	74.8	32	936	423	35	24	0.014	0.015	0.11	0.098	−12.6
At1g04400	77.8	32	3231	1707	90	78	0.013	0.012	0.18	0.272	−16.8
At1g04410	82.5	32	906	619	29	29	0.012	0.012	0.15	−0.014	−2.67
D) Region IV											
At5g03790	0	14	695	249	17	7	0.009	0.010	0.13	0.022	−1.89
At5g03810	8.6	14	1024	449	11	8	0.006	0.005	n.d.	−0.770	n.d.
At5g03830	13	14	987	664	46	36	0.018	0.022	0.11	1.29	−3.56
At5g03840 (*TFL1*)[f]	19.1	24	1023	621	8	6	0.003	0.0009	0.15	−2.06	0.848
At5g03840 promoter	20.1	24	636	650	34	34	0.015	0.021	0.14	1.55	1.25
At5g03850	22.9	25	609	489	10	10	0.005	0.004	0.14	−0.834	1.27
At5g03860	26.6	14	1162	413	10	8	0.006	0.006	0.21	−0.560	−0.923
At5g03870	30.2	14	1106	245	11	6	0.008	0.007	0.18	−0.273	−1.10
At5g03900	42.7	14	1048	601	5	5	0.003	0.002	0.12	−0.631	0.879

[a] S = number of nucleotide polymorphisms; [b] S_{sil} = number of silent nucleotide polymorphisms; [c] θ_W = estimated nucleotide diversity based on number of segregating sites; [d] π_{sil} = estimated pair-wise silent site diversity; [e] K_{JC} = Jukes-Cantor distance at silent sites between the ortholog in *A. lyrata;* [f] Focal low-diversity locus

n.a. = not applicable; n.d. = not determined

The focal loci in all these four regions have levels of silent site pair-wise nucleotide diversity (π_{sil}) that are over 10-fold lower than the genomic average of π_{sil} estimated to be between 0.006 and 0.009 (Nordborg et al. 2005; Schmid et al. 2005). Region I spans 113 kb and includes the low-diversity gene At1g67140, which encodes a polycystein cation domain channel protein and was identified as a "fast-evolving" gene in *A. thaliana* (Barrier et al. 2003). Both At1g67140 and the neighboring gene At1g67120 comprise much of this region, being 16.5 kb and 26 kb in length, respectively. We sequenced three 1 kb fragments across each of these relatively large genes (designated a, b, and c); At1g67140a is the original sequenced locus (Barrier et al. 2003).

Region II spans 76 kb and includes the recent gene duplicate At3g03700, which encodes a predicted sodium dicarboxylic acid symporter (Moore and Purugganan 2003). Region III spans 83 kb and includes At1g04300, which encodes a MATH domain protein and was identified as a low-diversity locus in a linkage disequilibrium study of the neighboring *CRYPTOCHROME2* (*CRY2*) locus (Olsen et al. 2004). Finally, region IV spans 43 kb and includes the focal gene At5g03840 (*TERMINAL FLOWER1* [*TFL1*]), which is involved in the developmental transition to flowering (Alvarez et al. 1992) and was initially analyzed in a population genetic analysis of flowering pathway genes (Olsen et al. 2002). Previous population genetic analyses of At1g67140 (region I), At3g03700 (region II), and the coding region of *TFL1*(region IV) suggest they are the targets of positive selection and have undergone selective sweeps (Barrier et al. 2003; Moore and Purugganan 2003; Olsen et al. 2002).

Two patterns of within-species nucleotide polymorphism are evident across the four genomic regions. Regions I and II typify one pattern; these two regions have broad valleys of reduced silent site nucleotide diversity (estimated by π_{sil} and θ_W) centered on the focal locus (Fig. 1A, 2A). In region I, we sequenced a total of 12,423 bps from nine genes and found a total of 137 polymorphic sites, 110 of which are silent site variants (Table 2). The focal gene At1g67140 is found in a broad 45 kb valley of reduced diversity that includes the neighboring gene, At1g67120, and is flanked by genes of moderately high diversity, At1g67110 and At1g67150. In region II, we sequenced a total of 9,823 bp from 11 loci and found 67 polymorphic sites, 36 of which are silent site changes (Table 2). Nucleotide diversity at the focal gene At3g03700 is at the lowest point of a shallow valley of low diversity spanning ~25 kb and including the neighboring loci At3g03690, At3g03710 and At3g03720. This valley is flanked on one side by a locus (At3g03680) with relatively elevated diversity.

The second pattern of nucleotide polymorphism is observed across regions III and IV. Focal low-diversity loci

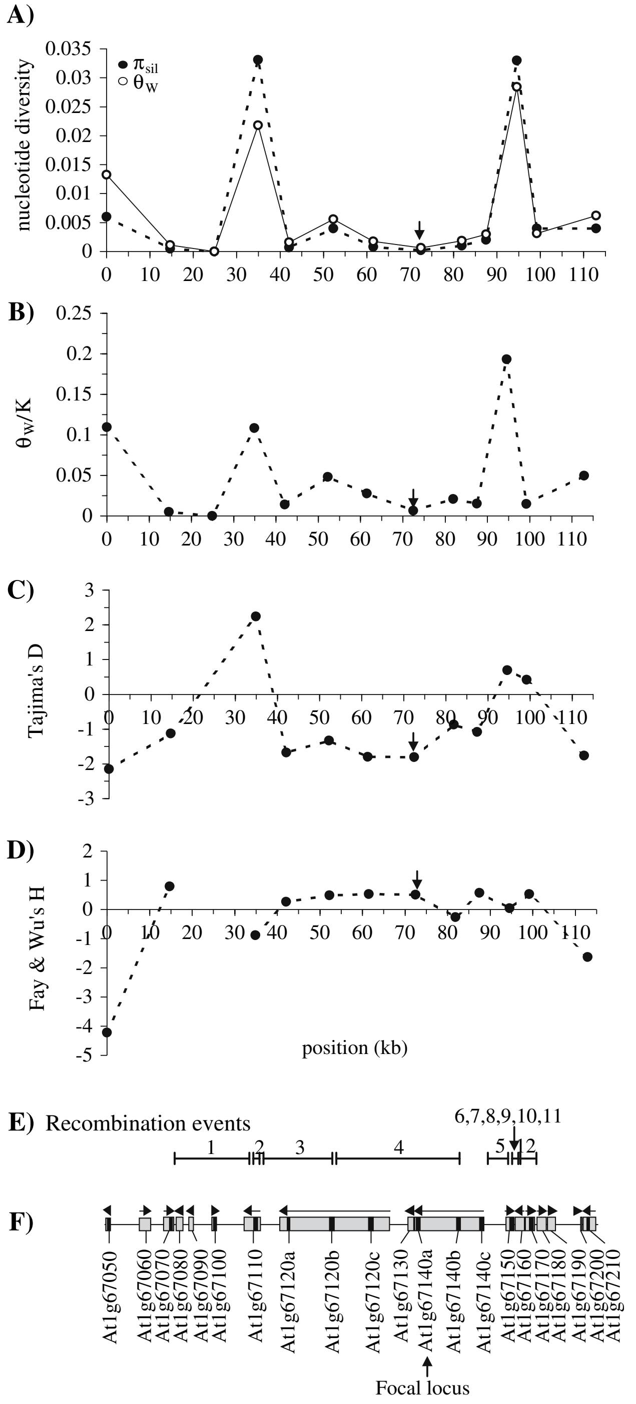

Fig. 1 The distribution of (**A**) nucleotide diversity (π_{sil} [closed circles] and θ_w [open circles]), (**B**) polymorphism to divergence ratio (θ_w/K), (**C**) Tajima's D, and (**D**) Fay and Wu's H in region I. (**E**) Position of predicted recombination events in region I (estimated from R_m). (**F**) Gene models for loci in region I (scaled to the positional axis of **A**–**D**). Horizontal arrows show transcriptional direction; genes are shaded grey and sequenced regions are shaded black (exon/intron structure not shown). A vertical arrow in **A**–**D** indicates the position of the focal locus At1g67140

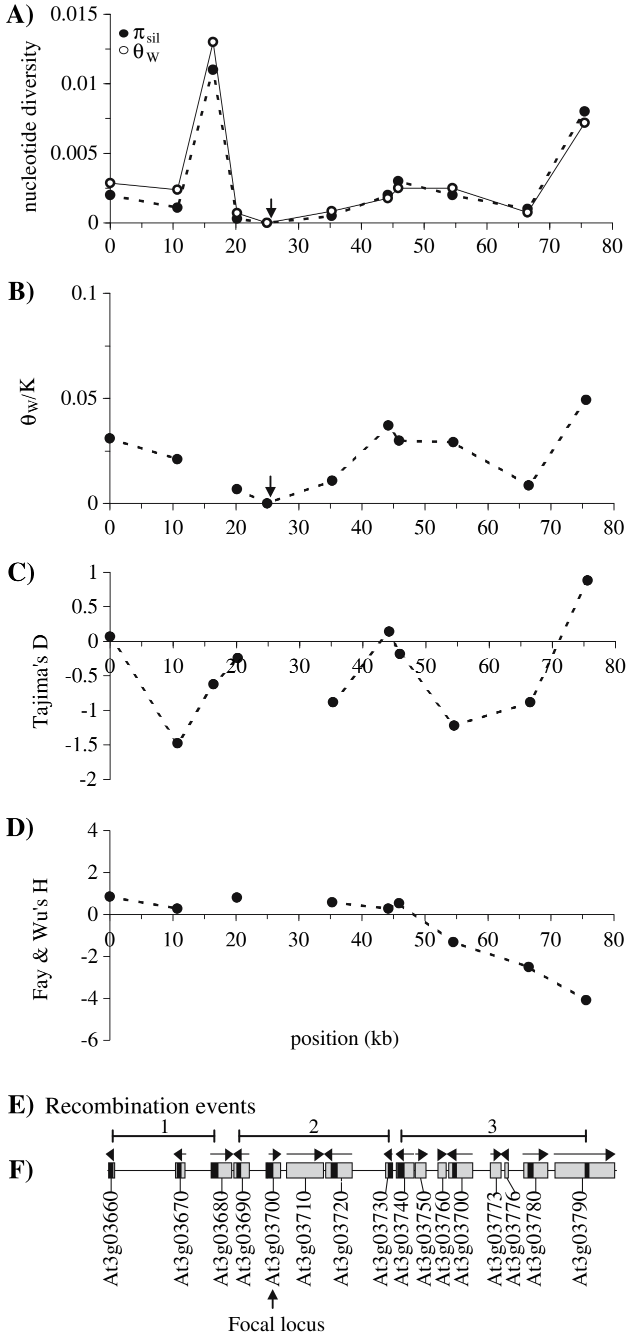

Fig. 2 The distribution of (**A**) nucleotide diversity (π_{sil} [closed circles] and θ_w [open circles]), (**B**) polymorphism to divergence ratio (θ_w/K), (**C**) Tajima's D, and (**D**) Fay and Wu's H in region II. (**E**) Position of predicted recombination events in region II (estimated from R_m). (**F**) Gene models for loci in region II (scaled to the positional axis of **A**–**D**). Horizontal arrows show transcriptional direction; genes are shaded grey and sequenced regions are shaded black (exon/intron structure not shown). A vertical arrow in **A**–**D** indicates the position of the focal locus At3g03700

in these regions are found in narrow valleys of reduced nucleotide diversity containing at most one other locus (Fig. 3A, 4A). In region III we sequenced a total of 17,295 bp from 18 loci and found 317 polymorphisms, 252 of which are silent substitutions (Table 2). Region III exhibits a number of alternating peaks and valleys of

Table 2 Average nucleotide diversity for each genomic region containing focal loci

Region	Region size (kb)	Total sites (bp)	Silent sites (bp)	S[a] total	S_{sil}[b] total	$\bar{\theta}_W$ ± s.e.[c] (V[d])	$\bar{\pi}_{sil}$ ± s.e. (V)	$\bar{K}_{JC}$ ± s.e. (V)
I	113642	12423	5717	137	110	0.007 ± 0.003 (1.3)	0.007 ± 0.003 (1.7)	0.13 ± 0.01 (0.35)
II	76351	9823	4621	67	36	0.003 ± 0.001 (0.84)	0.003 ± 0.001 (1.1)	0.10 ± 0.01 (0.32)
III	83446	17295	8458	317	252	0.006 ± 0.001 (0.80)	0.006 ± 0.001 (0.89)	0.14 ± 0.01 (0.26)
IV	43747	8290	4382	153	120	0.008 ± 0.002 (0.64)	0.009 ± 0.003 (0.90)	0.15 ± 0.01 (0.22)

[a] S = number of nucleotide polymorphisms; [b] S_{sil} = number of silent nucleotide polymorphisms; [c] s.e. = standard error; [d] V = coefficient of variation

nucleotide diversity. For example, in the 30 kb region between At1g04250 and At1g04350, π_{sil} changes 3- to 28-fold between neighboring loci. In region IV we sequenced a total of 8,290 bp from nine loci and found 153 polymorphic sites, 120 of which are silent substitutions (Table 2). Low levels of intraspecific polymorphism are restricted to the *TFL1* coding region, and nucleotide diversity increases 24-fold in the adjacent *TFL1* promoter and the neighboring locus, At5g03830.

Levels of silent-site polymorphism varied considerably within each genomic region. Differences between high and low π_{sil} values ranged from 24-fold (for region IV) to 165-fold (for region I), not including zero-diversity loci (Table 1). The coefficient of variation (V), a statistical measure of the degree of this variability, ranged from 0.9 to 1.7 for π_{sil} values and from 0.6 to 1.3 for θ_W (Table 2). This degree of variation in nucleotide polymorphism is on par with that found in a sampling of 876 loci randomly distributed across the *A. thaliana* genome for which V is 1.2 for π and 0.8 for θ_W (Nordborg et al. 2005). Changes in nucleotide diversity between adjacent loci were often abrupt and extreme in these regions. For example, π_{sil} values differed up to 47-fold between neighboring loci within each region (e.g., At1g67110 and At1g67150 in region I, At3g03680 in region II, At1g04290 in region III, and At5g03830 and the *TFL1* coding sequence in region IV; Figs. 1–4).

In order to assess the degree of similarity in nucleotide diversity levels between adjacent loci within each region, we calculated the average change in π_{sil} and θ_W between neighboring loci ($\Delta\pi_{sil}$ and $\Delta\theta_W$); (Table 3). The average $\Delta\pi_{sil}$ and $\Delta\theta_W$ values were lowest in region II ($\Delta\pi_{sil}$ and $\Delta\theta_W = 0.004$) and region III ($\Delta\pi_{sil}$ and $\Delta\theta_W = 0.005$) indicating less extreme changes in nucleotide diversity between neighboring loci and highest in region I ($\Delta\pi_{sil} = 0.011$ and $\Delta\theta_W = 0.009$) and region IV ($\Delta\pi_{sil} = 0.011$ and $\Delta\theta_W = 0.007$), which exhibit relatively large differences in nucleotide diversity at loci At1g67110 and At1g67150 in region I and between *TFL1* and neighboring loci in region IV. To analyze the significance of these values, we first calculated the mean $\Delta\pi_{sil}$ and $\Delta\theta_W$ values of 10,000 random arrangements of π_{sil} and θ_W values in each region (see Materials and Methods; Supplementary Fig. 1A–H). If there is a correlation in nucleotide diversity between neighboring loci, we should find less than 5% of the random arrangements with lower mean $\Delta\pi_{sil}$ or $\Delta\theta_W$ than the observed value. However, between 16% and 95% of the random arrangements had lower mean $\Delta\pi_{sil}$ and $\Delta\theta_W$ values than the observed value for each region, suggesting that nucleotide diversity estimates are uncorrelated between neighboring loci in these regions (Table 3).

Significant Variation in Effective Population Size (N_e), as Measured by θ/K, Exist Between Neighboring Loci

Variation in the levels of intraspecific polymorphism between loci may reflect differences in the underlying neutral mutation rate at those loci and not differences in selection regimes. It is more useful to track changes in the effective population size N_e across loci, as N_e should be equivalent between loci in the absence of selective forces. N_e can be approximated by the ratio of the estimate of the population mutation parameter ($\theta_W = 4\ N_e\mu$) to the intraspecific divergence ($K = 2\ T\mu$ under neutrality [Li 1997], where T is the divergence time between *A. thaliana* and *A. lyrata*); therefore, we tracked changes in this ratio across each region (Figs. 1–4B).

There is less variation in interspecific silent site divergence (K) between and within regions than in levels of intraspecific diversity, as measured by π_{sil} or θ_W (Table 2). Average K values for the regions range between 0.10 and 0.15 and are similar to the average of 0.14 substitutions/silent site determined from a genomic survey of intraspecific divergence between *A. thaliana* and *A. lyrata* (Schmid et al. 2005). Furthermore, V for the average divergence for each of the regions is three-to five-fold lower than the V for the average levels of intraspecific polymorphism (Table 3), suggesting that loci in these regions are diverging between the sister species at similar rates. Not surprisingly, the same general patterns in the distribution of intraspecific diversity also appear in the distribution of θ_W/K; that is, regions I

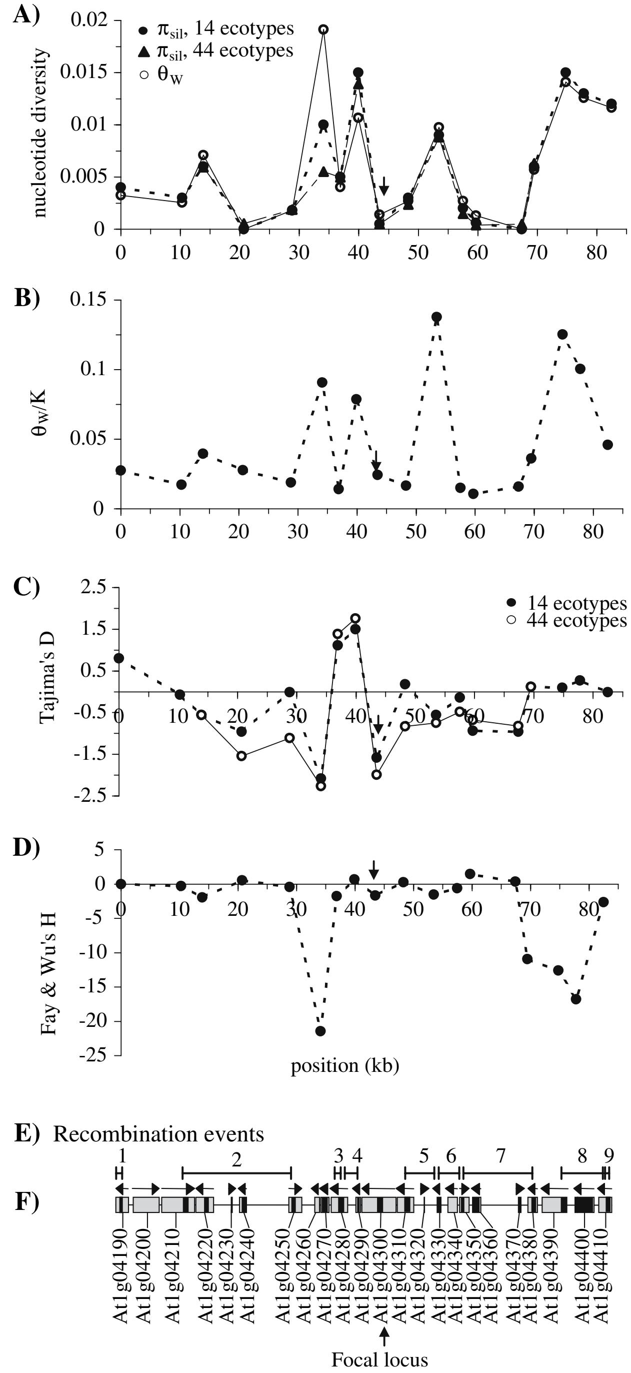

Fig. 3 The distribution of (**A**) nucleotide diversity (π_{sil} [closed circles] and θ_{w} [open circles]), (**B**) polymorphism to divergence ratio (θ_{w}/K), (**C**) Tajima's D, and (**D**) Fay and Wu's H in region III. For loci in the region spanning At1g04240 to At1g04360, estimates of π_{sil} (**A**) and Tajima's D (**C**) calculated from the base set of 14 ecotypes (closed circles) or and an expanded set of 44 ecotypes (closed triangles in A; open circles in D) are indicated. (**E**) Position of predicted recombination events in region III (estimated from R_m). (**F**) Gene models for loci in region III (scaled to the positional axis of **A-D**). Horizontal arrows show transcriptional direction; genes are shaded grey and sequenced regions are shaded black (exon/intron structure not shown). A vertical arrow in **A-D** indicates the position of the focal locus At1g04300

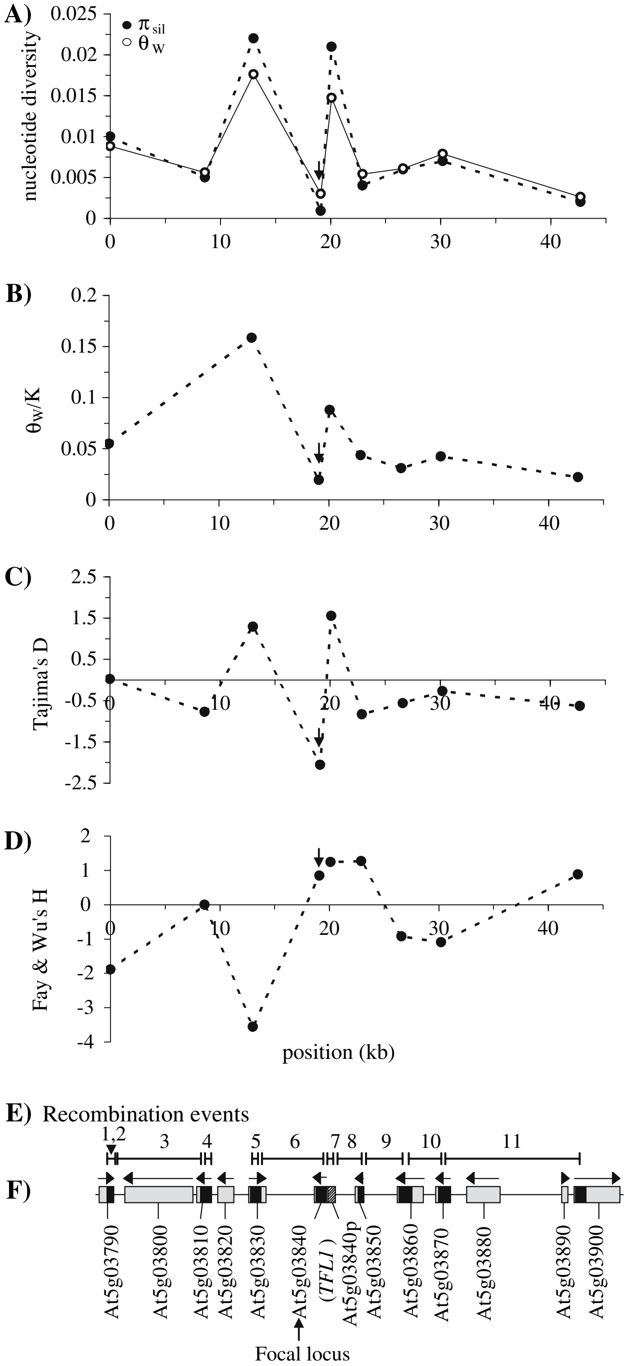

Fig. 4 The distribution of (**A**) nucleotide diversity (π_{sil} [closed circles] and θ_{w} [open circles]), (**B**) polymorphism to divergence ratio (θ_{w}/K), (**C**) Tajima's D, and (**D**) Fay and Wu's H in region IV. (**E**) Position of predicted recombination events in region IV (estimated from R_m). (**F**) Gene models for loci in region IV (scaled to the positional axis of **A–D**). Horizontal arrows show transcriptional direction; genes are shaded grey and sequenced regions are shaded black (exon/intron structure not shown). A vertical arrow in **A–D** indicates the position of the focal locus At5g03840 (*TFL1*)

and II exhibit broad valleys of reduced θ_{W}/K values, while regions III and IV exhibit alternating peaks and valleys of θ_{W}/K values (Figs. 1–4B).

Table 3 Average similarity in population genetic statistics (π_{sil}, θ_W, θ/K, Tajima's D, and Fay and Wu's H) between neighboring loci and mean distance between loci for each region

Region	mean distance between loci (kb)	mean $\Delta\pi_{sil}$ (P)	mean $\Delta\theta_W$ (P)	mean $\Delta(\theta_W/K)$ (P)	mean ΔD (P)	mean ΔH (P)
I	9.4	0.011 (0.65)	0.009 (0.73)	0.065 (0.92)	1.3 (0.34)	1.2 (0.18)
II	7.7	0.004 (0.40)	0.004 (0.42)	0.002 (0.17)	0.88 (0.58)	0.81 (0.002)
III	4.9	0.005 (0.16)	0.005 (0.25)	0.036 (0.52)	0.92 (0.04)	5.3 (0.12)
IV	5.3	0.011 (0.94)	0.007 (0.95)	0.061 (0.72)	1.6 (0.98)	1.5 (0.12)

P = proportion of 10,000 random arrangements of π_{sil}, θ_W,θ_W/K, D, or H values in each region with a lower mean than the observed value

For each region, we also calculated the mean change in θ_W/K between neighboring loci ($\Delta[\theta_W/K]$) as a measure of the degree of shared evolutionary history between loci. To test the correlation of θ_W/K between neighboring loci, we then compared the mean $\Delta(\theta_W/K)$ for each region to the mean(Δ θ_W/K) from 10,000 random arrangements of θ_W/K values in each region (Supplementary Fig. 1I–L). For regions I, III, and IV, between 52% and 92% of random arrangements had mean $\Delta(\theta_W/K)$ lower than the observed means (Table 3). For region II, we did not obtain divergence data for the highly polymorphic locus At3g03810, leading to more concordance in θ_W/K between neighboring loci. However, even without this locus, 17% of mean ($\Delta\theta_W$/K) values from the randomized dataset were lower than the observed mean (Table 3). Thus, N_e, as approximated by θ_W/K, varies significantly between neighboring loci in all of the regions we analyzed.

Tajima's D Values and Fay and Wu's H Values Vary Significantly Between Adjacent Loci in Most Genomic Regions

Because we observed such dramatic changes in diversity measures between linked loci, we were also interested in assessing the degree to which the polymorphism frequency distribution varied. Deviations in the polymorphism frequency spectrum from the neutral expectation can be tested using Tajima's D test statistic, while deviations in the frequency spectrum of derived polymorphisms can be tested with Fay and Wu's H statistic. Values of D and H are skewed toward the negative in regions in and around targets of recent selective sweeps (Fay and Wu 2000; Tajima 1989). We therefore determined the values of Tajima's D and Fay and Wu's H at loci in each region and assessed the degree to which these values were correlated between neighboring loci.

In region I, negative Tajima's D values span an approximate 45 kb region containing the focal locus At1g67140 and the neighboring loci At1g67130 and At1g67120 (Fig. 1C). This region is flanked by two loci, At1g67110 and At1g67150, with positive Tajima's D values. We could not calculate Tajima's D for the region II focal locus, At3g03700, which has no polymorphic sites. However, Tajima's D is negatively skewed across a 35 kb region surrounding At3g03700 (Fig. 2C). In contrast, focal loci in region III and region IV are found in very narrow valleys of negative Tajima's D values. Regions I, II, and IV exhibit significant variation in Tajima's D values between neighboring loci; the observed mean ΔD values for all four regions were greater than 5% of the mean ΔD of 10,000 randomized rearrangements of the data (Supplementary Fig. 5M–P, Table 3). In contrast, the mean ΔD value of region III was significantly lower ($P < 0.05$) than the randomized datasets.

All of the focal loci are flanked by loci with negative Fay and Wu's H, suggesting increased frequencies of derived polymorphism (Figs. 1–4D). As with Tajima's D values, the frequency spectrum of derived polymorphisms, as measured by Fay and Wu's H, was dissimilar between neighboring loci for the majority of the regions (Supplementary Fig. 1Q–T, Table 3). The exception was region II, where only 0.2% of the mean ΔH values from randomized data sets were lower than the observed value suggesting a significant correlation in Fay and Wu's H statistic between neighboring loci in this region.

Does Selection Play a Role in Shaping Patterns of Nucleotide Variation in the Regions?

Previous population genetic analyses identified three of the four focal loci (in regions I, II, and IV) as the targets of positive selection. A recent selective sweep at these loci would predictably result in reduced polymorphism and an excess of rare alleles and of high frequency derived alleles at linked loci. We used the CLR test of Kim and Stephan (2002) to determine if the patterns of nucleotide variation in each region were consistent with selection or neutral drift. For those regions for which the CLR test supported the selection hypothesis we implemented a GOF test to distinguish between variation patterns resulting from selection versus demographic processes such as population subdivision (Jensen et al. 2005).

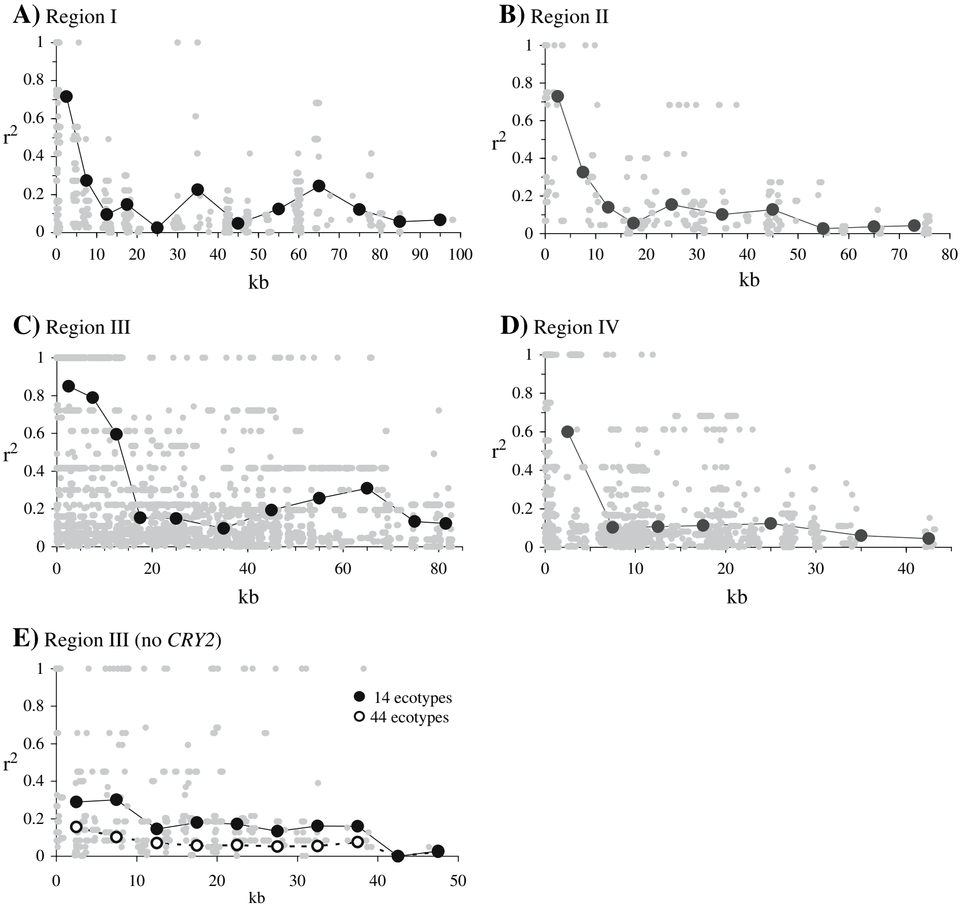

Fig. 5 LD (as measured by r^2) versus distance between polymorphic sites in (A) region I, (B) region II, (C) region III, and (D) region IV. (E) LD decay in region III for loci spanning At1g04240 to At1g04360 calculated from the base set of 14 ecotypes (closed circles) or and an expanded set of 44 ecotypes (open circles). Illustrating the decay in LD is a line that connects the average r^2 value measured for 5 kb intervals (for those points less than 20 kb distant) or 10 kb intervals (for those data values more than 20 kb distance)

Table 4 Composite-likelihood ratio (CLR) and goodness-of-fit (GOF) analysis for each region

Region	n	4N*r*[c] (per site)	Focal gene position (bp)	TEST A				TEST B			
				α[d]	X[e] (bp)	CLR1 (*P*)	GOF1 (*P*)	α[d]	X[e] (bp)	CLR2 (*P*)	GOF2 (*P*)
I	14	0.0002	72956	5.6	67110	7.444 (0.17)	NA	6.3	66396	10.73 (0.07)	NA
II	14	0.0003	25513	5.1	25793	4.88 (0.19)	NA	7.5	25575	7.51 (0.03)	1380 (0.57)
III	44	0.0005	23282	14.6	3552	14.8 (0.02)	3552 (0.001)	13.9	3114	9.04 (0.04)	15633 (0.001)
IV	14	0.0005	19628	5.4	40481	3.86 (0.31)	NA	10.6	40067	7.29 (0.18)	NA

[a] TEST A = derived alleles distinguished from ancestral alleles; [b] TEST B = derived alleles are not distinguished; [c] 4N*r* = scaled recombination rate per nucleotide; [d] α = estimate of the strength of selection (2N*s*); [e] X = estimated location of selected region

The CLR test supports the selection hypothesis for Regions II and III (Table 4). In Region II the CLR test was significant only when derived sites were not differentiated from ancestral sites (TEST B). This may be possible if the sweep was not recent, as simulations show that derived polymorphism will be lost after 0.4N$_e$ generations (Kim and Stephan, 2002). The GOF test supports the selection hypothesis suggesting that the patterns of variation are not due to demographic processes. Although the CLR test supports the selection hypothesis for Region III, the prediction location of the selective sweep is not centered on the focal locus for that region. Furthermore, unlike Region II, the GOF test statistic does not support the selection hypothesis. Given the extremely low *P* value for the GOF value for this region, it is probable that the significant CLR test result was due to underlying demography, such as population structure, and not selection (Jensen et al. 2005). Patterns of variation in Regions I and IV were indistinguishable from neutral simulations using the CLR test.

Recombination and the Extent of Linkage Disequilibrium

The fluctuating patterns of nucleotide diversity in genomic regions containing our focal low-diversity loci should be correlated with rates of recombination and the degree of LD at those regions, as recombination breaks down associations between neighboring loci. For example, we would predict that regions I and II, with broad valleys of reduced polymorphism, would be found in chromosomal regions with relatively low recombination and extended LD between loci. In contrast, we would expect regions III and IV, which have very narrow valleys of reduced polymorphism, to be found in chromosomal regions with elevated recombination rates and limited LD between loci. To assess these predictions, we estimated the global recombination rate at each region and analyzed the decay of LD within the regions (Fig. 5).

Both regions I and II are found in chromosomal positions with recombination estimates near the genome average of 4.8 cM/Mb (Zhang and Gaut 2003); the global recombination estimate is 4.9 cM/Mb for region I and 5.2 cM/Mb for region II. Region IV has a relatively elevated global recombination estimate of 7.8 cM/Mb, consistent with the extremely localized reduction in variability at the *TFL1* locus. However, region III, which exhibits a highly variable pattern of nucleotide diversity, has the lowest global recombination estimate of 3.0 cM/Mb.

Predictably, the rate of decay in LD parallels the trends in recombination estimates. In regions I and II, which exhibit broad valleys of reduced genetic diversity, LD, as measured by the square of the correlation coefficient (r^2), decays between 10–15 kb (Fig. 5A, B). In region IV, in which reduced diversity is limited to the focal *TFL1* gene, LD decays more rapidly, between 5–8 kb (Fig. 5D). In concordance with the low global estimate of recombination in the highly variable region III, LD in this region extends up to 15–20 kb (Fig. 5C). Furthermore, extensive LD ($r^2 = 1$) can be found between polymorphic loci up to 65 kb apart in region III. Thus, the fluctuations in nucleotide diversity between neighboring loci in regions I, II, and IV agree with global recombination estimates and the LD decay rate, while there is an apparent disjunction between these factors in region III. This may be due in part to the influence of the dimorphic *CRY2* locus in region III. If we focus on the 50 kb region from At1g04240 to At1g04370 that includes the focal gene At1g04300 but excludes *CRY2* and its flanking loci, LD decays rapidly, within 5 kb (Fig. 5E).

We also estimated Hudson's minimum number of recombination events (R_m) to see if differences in recombination rates reflect fluctuations in levels of nucleotide variation between linked loci. We found that R_m values correspond with the degree of fluctuation in nucleotide variation between neighboring loci in all regions except region I. For example, region II, with relatively low fluctuations in nucleotide variation, has an R_m of 3, whereas recombination is much more pervasive in regions III ($R_m = 9$) and IV ($R_m = 11$) which exhibit more frequent fluctuations in nucleotide variation. We mapped the recombination events in these regions and found that they potentially flank valleys of low diversity in each region (Figs. 1–4E). Region I, with the largest region of reduced variation, actually has an R_m of 12, on par with region IV. However, only three of these events occur in and around the 45 kb valley of reduced polymorphism surrounding the focal gene At1g67140 (recombination events #2–4, Fig. 1 E), while eight of these events map in or adjacent to the moderately high diversity flanking locus, At1g67150 (recombination events #5–12, Fig. 1E).

Discussion

Considerable Variation in Genetic Diversity Exists Between Neighboring Loci Within *A. thaliana* Genomic Regions

We found highly variable levels and patterns of nucleotide diversity in genomic regions surrounding low-diversity genes. This stands in contrast to other studies in other organisms which used autocorrelation analysis to find that neighboring loci have more similar polymorphism and divergence values (Hahn 2006; Matassi et al. 1999). We do not think this discrepancy is due a failure of our randomization method to detect correlation of diversity values between neighboring loci; our method also detects highly significant associations in polymorphism between linked regions on the Drosophila X chromosome ($P < 0.001$) when applied to the dataset of Hahn (2006). Our findings, however, do agree with previous analyses of local patterns of nucleotide polymorphism in *A. thaliana*. Similar fluctuations in genetic diversity have been reported for genomic regions surrounding other loci, such as *CLV2* (Shepard and Purugganan 2003), *MAM1* (Haubold et al. 2002), *FRI* (Hagenblad and Nordborg 2002; Hagenblad et al. 2004), and for some high diversity genes (Cork and Purugganan 2005). The ability to detect this variation in genetic diversity may be in part due to the relatively high density of markers surrounding the locus of interest used in this and other studies. This variation is still surprising, as linked loci should share similar evolutionary histories due to the low effective recombination rate in *A. thaliana* (Nordborg et al. 2005).

In studies of variation around the *FRI* locus, it has been suggested that population structure contributes to these fluctuations in population genetic statistics (Hagenblad and Nordborg 2002; Hagenblad et al. 2004). In an unstructured population, these values should be insensitive to sampling, but an increased sampling size of loci around *FRI* produced more uniform population genetics measurements, especially with respect to Tajima's D (Hagenblad et al. 2004). Note, however, that there was a two-fold decrease in marker density and more uniform marker distribution around the *FRI* locus in this later study which could also have contributed to a more uniform distribution of genetic variation (Hagenblad and Nordborg 2002; Hagenblad et al. 2004). The 14 *A. thaliana* accessions in our survey include members from across its native geographic range (Supplementary Table 1). To test for the underlying effects of population structure on the patterns of nucleotide polymorphism, we increased our sampling to 44 accessions for 10 loci surrounding At1g04300 in region III, one of the more variable regions. We found increased sampling only slightly altered estimates of nucleotide variation (π_{sil}), by 15% on average (see Fig. 3A). The effect of increased sampling on Tajima's D was more varied. The majority of loci showed only moderate adjustments to Tajima's D values, on the order of 26%, resulting in a smoothing of the distribution of Tajima's D estimates in the region (see Fig. 3C). However, upon further sampling, Tajima's D for two loci, At1g04310 and At1g04350, decreased more dramatically, by 3-to 5-fold, and Tajima's D for one locus, At1g04250, decreased over 83-fold, from −0.013 to −1.1 (see Fig. 3C). Thus, while population structure and sampling strategy may have contributed to fluctuation in certain population genetics statistics, such as Tajima's D, it did not appear to contribute strongly to the observed fluctuations in nucleotide variation.

These variations in nucleotide diversity are reflected in changes in $N_{e,}$ as measured by the θ_W/K ratio, suggesting discordance in evolutionary histories between neighboring loci in these genomic regions. It is possible that recombination has acted to break down associations between neighboring genes. The degree of dissimilarity in evolutionary histories between linked loci within a region generally correlated with estimates of recombination and linkage disequilibrium for that region, with the notable exception of region III. Nevertheless, recombination may still play a role in shaping diversity patterns in this highly variable region. While region III exhibits extensive LD and is found in a chromosomal region with low estimates of recombination (on par with centromeric regions), an explicit test of recombination estimates multiple recombination events across this region ($R_m = 9$). We also assessed LD decay in the 50 kb region surrounding the focal locus in our sample of 44 ecotypes in order to assess the influence of population structure on our estimates of LD. We observed an approximate 2-fold reduction in mean r^2 values when we increased our sampling to 44 ecotypes, although LD reached basal levels within 5 kb in both data sets (Fig. 5E). Thus our estimates of LD may be inflated due to underlying population structure in our original 14 sampled ecotypes.

It is also possible that gene conversion has contributed to the variation in nucleotide polymorphism levels exhibited in our regions. Gene conversion contributes significantly to the short-range pattern of LD in *A. thaliana* (Haubold et al. 2002; Nordborg et al. 2005). Indeed, recent estimates from genome-wide analyses suggest the gene conversion rate in *A. thaliana* is equal to the crossing over rate (Plagnol et al. 2006). Haubold et al. (2002) found that gene conversion rather than reciprocal recombination was the major contributing factor to the variation in nucleotide polymorphism in the 170 kb region around the *MAM1* gene. In this region 90% of recombination events were caused by gene conversion. Similarly, region III is found in a region of low recombination rate, yet the levels of polymorphism are not constant across the region; thus, gene conversion is a potential contributor to the observed pattern of polymorphism. Furthermore, studies in *Drosophila* have shown that gene conversion events associated with selective sweeps can give rise to peaks of high diversity within the sweep region (Glinka et al, 2006). This can potentially mask the signature of a selective sweep. We found a similar pattern in all three of the regions containing potentially selected loci (regions I, II, and IV). Each region has peaks of high diversity flanking the hypothesized selected locus; although, the CLR test supports the selection hypothesis only for region II.

Potential Contribution of Selection to the Evolutionary History of Low-Diversity Genes

Only three of our four low-diversity loci were previously predicted to be targets of positive selection: At1g67140 in region I, At3g03700 in region II and At5g03840 (*TFL1*) in region IV (Barrier et al. 2003; Moore and Purugganan 2003; Olsen et al. 2002). If selection is significantly strong and recent, this signature should extend to neighboring loci as predicted by the hitchhiking hypothesis (Kaplan et al. 1989; Kim and Stephan 2002; Przeworski 2002; Maynard Smith and Haigh 1974). This prediction is particularly important for distinguishing the pattern of selection from random fluctuations in nucleotide diversity due to drift (Kim and Stephan 2002; Przeworski 2002).

Patterns of nucleotide polymorphism are consistent with the hitchhiking prediction for only two of the loci, At1g67140 in region I and At3g03700 in region II, both of which are implicated from previous studies as targets of selection (Barrier et al. 2003; Moore and Purugganan 2003). Both loci are located in wide valleys of reduced polymorphism relative to divergence. For At1g67140, this valley spans approximately 45 kb and is flanked by two loci of elevated polymorphism relative to divergence. However, the CLR test of selection supports the selection hypothesis for only region II containing the recently duplicated gene, At3g03700. That the observed pattern of polymorphism in this region is a result of selection and not underlying population structure is bolstered by significantly low GOF test statistic. The predicted target of selection is centered on the region containing the focal locus, although the estimated strength of selection is low ($\alpha = 7.5$). Using an estimate of N_e of $4x10^5$ for *A. thaliana*, this equates to a selection coefficient (*s*) of approximately 10^{-5}, which is consistent with previous estimates of selection acting at this locus (Moore and Purugganan, 2003). It should be noted, however, that the power of the CLR to detect strong selection events is reduced in regions of low recombination, such as found in *A. thaliana* (Kim and Stephan, 2002).

Implications for the Future of Population Genetic Analyses in *A. thaliana*

These data have significant implications for inferring the evolutionary histories at individual loci in *A. thaliana*. Population genetic analyses of genomic regions, and not just individual genes, are necessary. For example, without analyzing the genomic context surrounding the gene of interest, we cannot rule out that selection is acting on neighboring loci. This is especially true for those loci with limited functional characterization or for which no prior reason for selection exists. And while this can sometimes lead to ambiguities as to the true target of selection, it can also strengthen the argument for the action of a particularly strong selection event in the region that includes the gene of interest. As genome-wide resequencing data for *A. thaliana* becomes available, such analyses will become the standard (Clark et al. 2007). Although researchers have also successfully used a "bottom-up" approach using genomic screens to identify potential targets of selection by scanning genomes for regions of extreme polymorphism levels, we must be careful when interpreting these results (Akey et al. 2002; Cork and Purugganan 2005; Diller et al. 2002; Mousset et al. 2003; Payseur et al. 2002; Vigouroux et al. 2002; Wiener et al. 2003). Ultimately, we will need to understand the functions of potentially selective genes in order to bolster their candidacy as targets of selection and to formulate adaptive hypotheses involving them.

Acknowledgments We thank Michael Purugganan and members of the Purugganan laboratory for stimulating discussions and thoughtful suggestions during the course of this research. We thank Ana Caicedo, Kitty Engelmann, and Kristie Mather for their critical reading of the manuscript. Furthermore, we thank Robert Noble and Michael Hughes of the Miami University department of mathematics and statistics for their statistical advice. Finally, we thank Yuseob Kim (Arizona State University) for his thoughtful advice on running the CLR analysis. This research was funded by an NIH postdoctoral fellowship to RCM and an NSF grant to MDP.

References

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

Alvarez J, Guli CL, Yu XH, Smyth DR (1992) *Terminal flower* - a gene affecting inflorescence development in *Arabidopsis thaliana*. Plant J 2:103–116

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Barrier M, Bustamante CD, Yu J, Purugganan MD (2003) Selection on rapidly evolving proteins in the *Arabidopsis* genome. Genetics 163:723–733

Barton NH (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355:1553–1562

Bloomfield P (2000) Fourier Analysis of Time Series. John Wiley & Sons, NY

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Scholkopf B, Nordborg M, Rätsch G, Ecker JR, Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317:338–342

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64

Cork JM, Purugganan MD (2005) High-diversity genes in the *Arabidopsis* genome. Genetics 170:1897–1911

Diller KC, Gilbert WA, Kocher TD (2002) Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. Mol Biol Evol 19:2342–2345

Diggle PJ (1990) Time Series: A Biostatistical Introduction. Oxford University Press, Oxford

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Glemin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc Biol Sci 273:3011–3019

Glinka S, De Lorenzo D, Stephan W (2006) Evidence of gene conversion associated with a selective sweep in *Drosophila melanogaster*. Mol Biol Evol 23:1869–1878

Hagenblad J, Nordborg M (2002) Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161:289–298

Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, Zheng H, Marjoram P, Weigel D, Nordborg M (2004) Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics 168:1627–1638

Hahn MW (2006) Accurate inference and estimation in population genomics. Mol Biol Evol 23:911–918

Haubold B, Kroymann J, Ratzka A, Mitchell-Olds T, Wiehe T (2002) Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. Genetics 161:1269–1278

Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. Genet Res 50:245–250

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolism. Academic Press, New York, p. 21–132

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–1410

Kaplan NL, Hudson RR, Langley CH (1989) The “hitchhiking effect” revisited. Genetics 123:887–899

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777

Li W-H (1997) Molecular Evolution. Sinauer Associates, Inc., Sunderland, MA

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant Journal 4:745–750

Matassi G, Sharp PM, Gautier C (1999) Chromosomal location effects on gene sequence evolution in mammals. Curr Biol 9:786–791

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 100:15682–15687

Mousset S, Brazier L, Cariou ML, Chartois F, Depaulis F, Veuille M (2003) Evidence of a high rate of selective sweeps in African *Drosophila melanogaster*. Genetics 163:599–609

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396:572–575

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of Arabidopsis *CRY2* flowering time alleles. Genetics 167:1361–1369

Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD (2002) Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics 160:1641–1650

Payseur BA, Cutter AD, Nachman MW (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol 19:1143–1153

Plagnol V, Padhukasahasram B, Wall JD, Marjoram P, Nordborg M (2006) Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in *Arabidopsis thaliana*. Genetics 172:2441–2448

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174–175

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa, NJ, p. 365–386

SAS Institute (1990) SAS/ETS User’s Guide ,Version 6. SAS Institute Inc., Cary, NC, USA

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T (2005) A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics 169:1601–1615

Shepard KA, Purugganan MD (2003) Molecular population genetics of the Arabidopsis *CLAVATA2* region: the genomic scale of variation and selection in a selfing species. Genetics 163:1083–1095

Shimizu KK, Cork JM, Caicedo AL, Mays CA, Moore RC, Olsen KM, Ruzsa S, Coop G, Bustamante CD, Awadalla P, Purugganan MD (2004) Darwinian selection on a selfing locus. Science 306:2081–2084

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tiffin P, Hacker R, Gaut BS (2004) Population genetic evidence for rapid changes in intraspecific diversity and allelic cycling of a specialist defense gene in *Zea*. Genetics 168:425–434

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci USA 99:9650–9655

Watterson GA (1975) Number of segregating sites in genetic models without recombination. Theoretical Population Biology 7:256–276

Wiener P, Burton D, Ajmone-Marsan P, Dunner S, Mommens G, Nijman IJ, Rodellar C, Valentini A, Williams JL (2003) Signatures of selection? Patterns of microsatellite diversity on a chromosome containing a selected locus. Heredity 90:350–358

Williams EJ, Hurst LD (2000) The proteins of linked genes evolve at similar rates. Nature 407:900–903

Zhang L, Gaut BS (2003) Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res 13:2533–2540