Molecular Population Genetics of the Arabidopsis *CLAVATA2* Region: The Genomic Scale of Variation and Selection in a Selfing Species

Kristen A. Shepard[1] and Michael D. Purugganan[2]

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695

Manuscript received July 16, 2002
Accepted for publication November 20, 2002

ABSTRACT

The *Arabidopsis thaliana CLAVATA2* (*CLV2*) gene encodes a leucine-rich repeat protein that regulates the development of the shoot meristem. The levels and patterns of nucleotide variation were assessed for *CLV2* and 10 flanking genes that together span a 40-kb region of chromosome I. A total of 296 out of 7959 sequenced nucleotide sites were polymorphic. The mean levels of sequence diversity of the contiguous genes in this region are approximately twofold higher than those of other typical Arabidopsis nuclear loci. There is, however, wide variation in the levels and patterns of sequence variation among the 11 linked genes in this region, and adjacent genes appear to be subject to contrasting evolutionary forces. *CLV2* has the highest levels of nucleotide variation in this region, a significant excess of intermediate frequency polymorphisms, and significant levels of intragenic linkage disequilibrium. Most alleles at *CLV2* are found in one of three haplotype groups of moderate (>15%) frequency. These features suggest that *CLV2* may harbor a balanced polymorphism.

BALANCED polymorphisms are maintained in populations by selective forces acting on alternative alleles of a locus (Richman 2000; Tian *et al.* 2002). Various forms of balancing selection as well as local adaptation can lead to the persistence of allelic variants of a gene in a species. Molecular population genetic analyses have identified several examples of balanced polymorphisms in eukaryotic genes, including the *Adh* locus in *Drosophila melanogaster* (Kreitman and Hudson 1991), the self-incompatibility locus in various plant species (Richman 2000; Uyenoyama 2000), and the *Rpm1* disease-resistance gene in *Arabidopsis thaliana* (Stahl *et al.* 1999). In balanced polymorphisms, selection is expected to maintain a region of enhanced variability of neutral polymorphisms surrounding a selected site, resulting in correlated gene genealogies among linked loci (Nordborg *et al.* 1996; Tian *et al.* 2002). The window of increased variation in outcrossing species, however, can be fairly narrow as recombination breaks apart correlations among linked sites surrounding a target of balancing selection (Nordborg *et al.* 1996). In general, the scale of elevated variation in species such as *D. melanogaster* is <1 kb; variation around the Fast/Slow *Adh* polymorphism, for example, is enhanced in a region of ~200 bp (Kreitman and Hudson 1991).

In selfing species, the width of the genomic region of enhanced variation scales with the inverse of the population recombination parameter $C = 4N_e r'$, where N_e is the effective population size and r' is the selfing-reduced effective recombination rate (Charlesworth *et al.* 1997). Since r' in selfing species is generally lower than the recombination rate in outcrossing species, the window of enhanced variation surrounding a balanced polymorphism should be wider in selfers than in outcrossers. Indeed, a very low recombination rate can result in balanced polymorphisms encompassing large tracts of linked sites in the genome. Thus, in selfing species, selection for balanced polymorphism can thus affect the genetic diversity and evolutionary dynamics of both adjacent and distant genes.

A. thaliana provides an excellent opportunity to empirically assess the genomic impact of balanced polymorphisms in a predominantly selfing plant species. Outcrossing rates in this weedy plant species are estimated to be as low as 1% (Abbot and Gomez 1989). Species-wide surveys of nucleotide variation reveal a low level of recombination within nuclear loci (Kawabe *et al.* 1997; Miyashita *et al.* 1998; Nordborg *et al.* 2002). This low effective recombination rate may lead to strong correlations among nucleotide polymorphisms over long distances in the genome. In *A. thaliana*, linkage disequilibrium among polymorphic nucleotide sites is observed both within and among genes, and disequilibrium tracts can extend up to ~250 kb (Nordborg *et al.* 2002). In contrast, linkage disequilibrium generally

Sequence data from this article have been deposited with the EMBL/GenBank Data libraries under accession nos. AF528566–AF528713.

[1]*Present address:* Department of Biological Sciences, Barnard College, Columbia University, New York, NY 10027.

[2]*Corresponding author:* Department of Genetics, 3513 Gardner Hall, North Carolina State University, Raleigh, NC 27695. E-mail: michaelp@unity.ncsu.edu

decays within several hundred basepairs in *D. melanogaster* (Long *et al.* 1998) and within 1.5 kb in the outcrossing plant *Zea mays* (Remington *et al.* 2001; Tenaillon *et al.* 2001).

Low effective recombination and long-range linkage disequilibrium in *A. thaliana* suggest that the region of enhanced variation associated with a balanced polymorphism could extend over several linked genes. This linkage may affect the rate and efficacy of selection on alternate alleles. Recent studies, however, contradict this prediction; the effects of balanced polymorphisms in the *TFL1* gene (Olsen *et al.* 2002), the *RPS5* disease-resistance locus (Tian *et al.* 2002), and the enzyme locus *PgiC* (Kawabe *et al.* 2000) appear to be quite localized. To clarify how far the effects of selection can extend in the *A. thaliana* genome, we have undertaken a systematic investigation of a putative balanced polymorphism in the *CLAVATA2* (*CLV2*) gene.

CLV2 is a meristem regulatory gene located near 89 cM on chromosome I. Loss-of-function mutations at *CLV2* result in the accumulation of undifferentiated cells in vegetative, inflorescence, and floral meristems. The enlargement of these shoot meristems contributes to the formation of extra flowers and floral organs (Kayes and Clark 1998). The 720-amino-acid (aa) protein encoded by *CLV2* includes a signal peptide, a putative extracellular domain with ~20 leucine-rich repeats (LRR), a transmembrane region, and a short cytoplasmic domain (Jeong *et al.* 1999). Although CLV2 is structurally similar to the Cf family of disease-resistance proteins from tomato, the complexes formed by CLV2 and the Cf proteins are quite different (Rivas *et al.* 2002). CLV2 appears to be necessary for protein accumulation of CLV1, a LRR receptor kinase (Jeong *et al.* 1999). These two proteins are hypothesized to form a disulfide-linked heterodimer in the plasma membrane, although there is not yet direct evidence for this interaction. When bound by a multimeric ligand that includes the CLV3 protein (Trotochaud *et al.* 2000), the activated CLV1-CLV2 complex triggers a signal transduction cascade that ultimately represses the *WUSCHEL* gene, a transcription factor gene that promotes shoot meristem growth (Trotochaud *et al.* 1999; Brand *et al.* 2000).

The initial isolation of *CLV2* showed that this gene harbors a large amount of nucleotide diversity (Jeong *et al.* 1999). This elevated variation might be caused by the maintenance of a balanced polymorphism in *CLV2*, by correlated effects of selection at neighboring loci, or by a high rate of mutation in this portion of the genome. Here we report a molecular population genetic analysis of *CLV2* and 10 adjacent genes that are found in a 40-kb region of the genome. Consistent with the low effective recombination in *A. thaliana*, linkage disequilibrium persists not only within genes in the *CLV2* region, but also between loci separated by as much as 25 kb. Genes in this region also show the elevated silent site nucleotide variation associated with the effects of balancing selection. The level of variation is highest at *CLV2*; however, there is a nearly 10-fold range in the levels of nucleotide diversity among the neighboring loci. Moreover, the allelic distribution of nucleotide variation varies markedly in this region. Although *CLV2* gene also has a significant excess of intermediate frequency polymorphisms and intragenic linkage disequilibrium (consistent with balancing selection), most nearby loci have an excess of rare polymorphisms. These results indicate that adjacent genes may have differing patterns and levels of nucleotide variation, suggesting that they are subject to contrasting evolutionary forces.

MATERIALS AND METHODS

Isolation and sequencing of alleles: *A. thaliana* ecotypes were obtained from single-seed propagated material provided by the *Arabidopsis* Biological Resource Center (ABRC; see Table 1). The Lisse-2 seed stock was from the population collection of P. H. Williams maintained at ABRC. *A. lyrata* seed from a Karhumaki, Russia, population was provided by O. Savolainen and Helmi Kuittinen.

Genomic DNA was isolated from young leaves of 9–20 *A. thaliana* ecotypes and one *A. lyrata* accession using the plant DNeasy mini kit (QIAGEN, Chatsworth, CA). PCR primers for 11 genes in this region were designed from the Col-0 genomic sequence [bacterial artificial chromosome (BAC) T8F5, GenBank accession no. AC004512] using Primer3 (Rozen and Skaletsky 1998); all primers were located in predicted exons. Primers were chosen without regard to predicted functional domains, but are biased toward the 5′ end of coding sequences. Description of sequenced genes (see Table 2) and the PCR primers used in the amplification reactions are described in Supplementary Text I at http://www.genetics.org/supplemental/. PCR of *A. thaliana* samples was performed with *Taq* DNA polymerase (Eppendorf, Madison, WI) using protocols designed for direct sequencing. PCR of *A. lyrata* samples was performed with the error-correcting *Pwo* polymerase (Roche) using the manufacturer's amplification protocol. The error rate of this error-correcting polymerase is <1 in 7000 bp (M. Purugganan, unpublished observations).

DNA fragments were purified using the QIAquick PCR purification kit or the QIAquick gel extraction kit (QIAGEN). *A. thaliana* samples were sequenced directly via cycle sequencing with BigDye terminators (Applied Biosystems) using the primers described in Supplementary Text I at http://www.genetics.org/supplemental/. Several singleton polymorphisms were confirmed with reamplification and sequencing. Amplified *A. lyrata* products were cloned into pCR4Blunt-TOPO vector using the Zero Blunt TOPO PCR cloning kit (Invitrogen). Plasmid miniprep DNA was isolated using the QIAprep miniprep kit (QIAGEN), and sequenced twice via cycle sequencing from both directions. DNA sequencing was conducted with a Prism 3700 96-capillary automated sequencer (Applied Biosystems). The PHRED and PHRAP functions (Ewing and Green 1998; Ewing *et al.* 1998) of BioLign 2.0.7 (Tom Hall, North Carolina State University) were used to call bases and to create contigs; low-quality sequence was trimmed from contigs. GenBank accession numbers for these genes are AF528566–AF528713.

Molecular population genetic data analysis: Sequences used in this study were visually aligned against the *A. thaliana* GenBank sequence for the Col-0 accession (no. AC004512). The variable length portions of microsatellites were excluded from

the analysis. The *A. lyrata* ortholog was used as the outgroup. Interspecific divergence distances were estimated from silent sites with the Kimura two-parameter model using MEGA2.1 (Kumar *et al.* 2001). Polymorphism analyses were conducted using DnaSP 3.51 (Rozas and Rozas 1999). Levels of nucleotide diversity per site were estimated as π (Tajima 1983) and θ_W (Watterson 1975). The Tajima (1989) and Fu and Li (1993) tests for selection were conducted; Fu and Li's test was performed both with (*D*) and without (*D**) the *A. lyrata* outgroup sequence. Significance of Tajima's and Fu and Li's test statistics was determined in coalescent simulations with 10,000 runs using the number of segregating sites under a model of no recombination. Linkage disequilibrium between informative sites within and between genes was estimated as r^2 (Hill and Robertson 1968) with significance determined by Fisher's exact tests. Levels of intragenic disequilibrium were also quantified by the Z_{nS} statistic (Kelly 1997) with deviation from neutral-equilibrium expectations determined by coalescent simulations with 10,000 runs using the recombination parameter estimated from the data.

The Hudson-Kreitman-Aguadé (HKA) two-locus test (Hudson *et al.* 1987) was conducted using silent site changes from a program available from Jody Hey (Rutgers University). The *Adh* locus was chosen as the reference neutral locus in these tests (Innan *et al.* 1996; Miyashita *et al.* 1998). Some studies suggest that this gene may harbor a balanced polymorphism (Miyashita 2001), which may indicate that using this gene as a reference locus is conservative when testing for the hypothesis of balancing selection. Among several genes, however, the pattern of variation at *Adh* is one that is most consistent with neutral-equilibrium expectations under a metapopulation model (Innan and Stephan 2000).

Previously published *A. thaliana* sequences of the following genes, which were available at the time of this study, were used in comparisons of nucleotide diversity: *Adh* (Innan *et al.* 1996; Miyashita *et al.* 1998); *AP1*, *LFY*, and *TFL1* (Olsen *et al.* 2002); *AP3* and *PI* (Purugganan and Suddith 1999); *CAL* (Purugganan and Suddith 1998); *CHI* (Kuittinen and Aguadé 2000); *ChiA* (Kawabe *et al.* 1997); *ChiB* (Kawabe and Miyashita 1999); *F3H* and *FAH1* (Aguadé 2001); *PgiC* (Kawabe *et al.* 2000); and *RPS2* (Caicedo *et al.* 1999). The same set of genes, with the exception of *ChiB* and *RPS2*, was used in interspecific divergence comparisons.

TABLE 1

***A. thaliana* accessions**

Ecotype designation	Locality	ABRC seed accession
	Sampled for all loci	
Bs-1	Basel, Switzerland	CS996
Bla-1	Blanes/Gerona, Spain	CS970
Col-0	Landsberg, Germany	—
Chi-1	Chisdra, Russia	CS1074
Co-1	Coimbra, Portugal	CS1084
Cvi-0	Cape Verde Islands	CS902
Gr-3	Graz, Austria	CS1202
Kas-1	Kashmir, India	CS903
Ita-0	Ibel Tazekka, Morocco	CS1244
Lisse-2[a]	Lisse, Netherlands	CS6092
	Sampled only at CLV2 and the three nearest loci	
An-2	Antwerpen, Belgium	CS6604
Bu-0	Burghaun/Rhon, Germany	CS6632
El-0	Ellershausen, Germany	CS6694
Fi-0	Frickhofen, Germany	CS6704
JI-3	Vranov u Brna, Czechoslovakia	CS5745
Lan-0	Lanark, Great Britain	CS6768
Ler-0	Landsberg, Germany	CS20
Lu-1	Lund, Sweden	CS1352
Pi-0	Pitzal/Tirol, Austria	CS6832
Sf-1	San Feliu, Spain	CS6855
Ws-0	Wassileskija, Ukraine	CS915

[a] Field strain from the Williams collection (ABRC). The name is not an official ecotype designation.

RESULTS

Nucleotide variation among linked genes in a 40-kb region of Arabidopsis chromosome I: Fragments of 11 adjacent genes on chromosome I were sequenced from 9 to 12 *A. thaliana* accessions sampled primarily from Eurasia (Tables 1 and 2). The sequenced regions spanned exons and (when present) introns within the coding region of each gene; fragments ranged from 277 to 939 bp, with a mean length of 724 bp/gene. Of the 7959 nucleotide sites sequenced for this study, 296 sites segregated for single nucleotide polymorphisms. Twenty-eight indel polymorphisms, ranging from 1 to 3.9 kb, were also observed in these sequences. Four indels, two in the serpin and two in the ARI/RING-like gene, are associated with simple sequence or microsatellite repeats in introns. Seven indels occur in coding regions. Tables of polymorphic sites are given in Supplementary Figures S1–S7 at http://www.genetics.org/supplemental/.

Polymorphisms in the *UBQ13*, the MATH domain gene, and the serpin suggest that these loci may be pseudogenes. All sampled *UBQ13* alleles contain a partial ubiquitin repeat followed by three or four complete repeats. We were unable to locate the rest of the repeat in the upstream genomic sequence of the Col-0 accession. The internal repeats appear to have undergone substantial recombination; because homology among these repeats was difficult to determine, analyses were restricted to the 5′ flanking region, the partial repeat, and the first and last complete repeats. One allele of *UBQ13* codes for a premature stop codon, while two alleles contain 3- or 12-bp deletions in coding sequence. The Col-0 allele contains a 3.9-kb insertion of mitochondrial DNA (Sun and Callis 1993) that was not observed in any other accession. One allele of the MATH domain gene codes for a premature stop codon, while another allele has a frameshift mutation. The putative serpin gene has multiple lesions, including three alleles with premature stop codons, three with frameshift mutations, and five with a 39-bp deletion in the coding region. The large number of potentially deleterious polymorphisms in the *UBQ13* and serpin genes suggests that they are recent pseudogenes. It is unclear whether the MATH domain gene, which is expressed in Col-0, segregates for rare deleterious alleles or is an incipient pseudogene. In estimating levels of silent site nucleotide

TABLE 2

Genes in the *A. thaliana CLV2* genomic region

Gene	Gene ID[a]	Position[b] (kb)
AGL37	AT1G65330	32.66–33.39
CYP96A3	AT1G65340	34.51–35.53
UBQ13[c]	AT1G65345[d]	38.52–43.47
AGL23	AT1G65360	47.48–48.17
MATH domain	AT1G65370	50.78–51.75
CLAVATA2	AT1G65380	53.16–54.23
serpin ϕ	AT1G65385	56.17–57.03
TIR domain	AT1G65390[d]	58.63–59.40
AtNAP11	AT1G65410	62.01–63.03
Antigen receptor	AT1G65420	64.07–63.73
ARI/RING-like	AT1G65430	71.11–72.12

[a] Based on The Arabidopsis Information Resource genome annotation.
[b] Primer positions along sequence of BAC clone T8F5.
[c] Based only on partial, first, and last repeat units.
[d] Complete gene combines two gene IDs (see Supplementary Text); ϕ pseudogene.

diversity for these loci, we have aligned these genes according to their predicted coding potential.

Estimates of silent nucleotide diversity and divergence in the *CLV2* region: Nucleotide diversity at silent sites (third position of codons and noncoding regions) for these 11 genes was estimated from the average number of pairwise differences (π; Tajima 1983) and from the number of segregating sites (θ_W; Watterson 1975). Focusing on silent sites permits comparisons of sequences with different proportions of coding to noncoding sequences. In addition, the amount of silent site diversity provides information about the action of selection at linked sites. Among the genes in the *CLV2* region, levels of π span nearly one order of magnitude, from 0.0063 to 0.0579 (see Table 3A and Figure 1). Levels of θ_W show a comparable range, from 0.0075 to 0.0489 (Table 3A). *CLV2* exhibits the greatest silent site diversity, with the highest π and the second highest θ. The high value of θ_W observed at the putative antigen receptor can be attributed to a single allele from the Ita-0 accession that accounts for 7 of the 10 segregating sites in this gene. The lowest levels of π and θ_W were observed 3 kb upstream of *CLV2* in the MATH domain gene and 17 kb downstream in the ARI/RING-like gene.

Overall, the *CLV2* region exhibits elevated levels of silent site nucleotide diversity compared to other nuclear genes in *A. thaliana*. The mean values of π and θ_W for the 11 genes in the *CLV2* region are 0.0219 ± 0.005 and 0.0241 ± 0.004, respectively. These mean diversity levels are considerably greater than those observed among 14 previously studied *A. thaliana* genes. For these other loci (see materials and methods), the mean values of silent site π and θ_W are 0.009 ± 0.001 and 0.012 ± 0.002, twofold lower than those of the genes in the *CLV2* region (see Figure 2). In contrast, the 11 genes in the *CLV2* region display only slightly higher levels of nucleotide divergence between *A. thaliana* and the closely related species *A. lyrata* (Table 3A). The mean level of silent site sequence divergence, K_{2P}, between these two species for 12 previously studied *A. thaliana* genes is 0.123 ± 0.007 substitutions/site. The mean nucleotide divergence level for the 11 linked genes in the *CLV2* region is 0.138 ± 0.010 substitutions/site.

The HKA test (Hudson *et al.* 1987) detects differences in nucleotide variation levels between two loci when corrected for mutation rate variation. This test was applied to the genes in the *CLV2* region, with the *Adh* gene serving as the reference locus. The observed numbers of silent intraspecific polymorphisms and interspecific differences for *Adh* are 30 and 124.12, respectively. The numbers of silent site within-species polymorphisms and between-species differences for each gene at the *CLV2* region are indicated in Table 3. The HKA tests reveal that three genes have significant increases in nucleotide variation levels (Table 4A). The three genes that display significant deviation from the neutral-equilibrium model based on the HKA test are *CLV2* ($P < 0.01$), *AtNAP11* ($P < 0.03$), and the antigen receptor gene ($P < 0.04$). The non-neutral evolution at these loci is associated with an excess of intraspecific variation for each gene as compared to the neutral *Adh* locus.

Selective forces among linked genes in the *CLV2* region: The frequency distribution of polymorphisms provides information on the relative roles of neutral drift *vs.* selection at specific loci. The skewness of frequency distributions for nucleotide polymorphisms in the sample or along branches in the gene genealogy can be evaluated with the Tajima (Tajima 1989) or Fu and Li (Fu and Li 1993) tests for selection, respectively. Since *A. thaliana* may have experienced a recent population expansion, these two tests should be interpreted with caution when inferring selection. However, they may still provide information on the extent and direction of deviations in molecular diversity patterns from predictions of the neutral-equilibrium model, as well as permit comparison of relative patterns of nucleotide variation between genes. To take into account the selfing nature of *A. thaliana*, the significance of these test statistics was assessed by coalescent simulations under a stringent model of no recombination.

Among the 11 genes in the *CLV2* region, 8 have negative values of Tajima's *D* and Fu and Li's *D* and *D**, indicating an excess of low-frequency polymorphisms within these loci (Table 4A). The trend toward excess low-frequency polymorphism for most of the genes at the *CLV2* region is similar to that observed for many other Arabidopsis nuclear genes (Purugganan and Suddith 1999; Innan and Stephan 2000; Kuittinen and Aguadé 2000). This pattern of variation may reflect the inbreeding associated with this selfing plant and/

TABLE 3

Features of sequence variation in the *A. thaliana CLV2* genomic region

Gene	n^a	Length[b] (kb)	No. of silent sites	S[c]	S (silent)[d]	π^e	$\theta_W{}^e$	$K_{2P}{}^f$
A. Initial sampling at all loci								
AGL37	10	652	142.8	8	4	0.0116 ± 0.007	0.0099 ± 0.0060	0.138 ± 0.027 (23.25)
CYP96A3	9	939	203.7	33	17	0.0327 ± 0.0183	0.0307 ± 0.0144	0.118 ± 0.019 (31.89)
UBQ13	10	543	149.6	16	6	0.0124 ± 0.0097	0.0142 ± 0.0078	0.168 ± 0.039 (19.95)
AGL23	9	636	238.5	32	13	0.0156 ± 0.0123	0.0201 ± 0.0097	0.188 ± 0.028 (36.94)
MATH domain	9	837	425.5	15	9	0.0063 ± 0.0050	0.0078 ± 0.0040	ND
CLAVATA2	12	924	223.4	50	31	0.0579 ± 0.0254	0.0460 ± 0.0191	0.119 ± 0.018 (35.38)
serpin ϕ	10	723	353.0	38	25	0.0205 ± 0.0143	0.0250 ± 0.0111	0.155 ± 0.022 (45.95)
TIR domain	10	673	232.2	39	20	0.0288 ± 0.0177	0.0305 ± 0.0137	0.151 ± 0.023 (37.30)
AtNAP11	10	885	604.4	43	43	0.0191 ± 0.0139	0.0252 ± 0.0107	0.097 ± 0.012 (56.70)
Antigen receptor	10	277	65.1	10	9	0.0294 ± 0.0312	0.0489 ± 0.0247	0.149 ± 0.041 (13.10)
ARI/RING-like	10	870	614.5	14	13	0.0069 ± 0.0045	0.0075 ± 0.0036	0.094 ± 0.012 (53.50)
B. After additional sampling at *CLV2* and the three nearest loci								
MATH domain	19	801	419.5	20	13	0.0060 ± 0.0052	0.0087 ± 0.0038	ND
CLAVATA2	21	934	225.2	50	31	0.0558 ± 0.0206	0.0383 ± 0.0142	0.116 ± 0.018 (34.87)
serpin ϕ	20	731	353.7	43	29	0.0192 ± 0.0125	0.0231 ± 0.0087	0.157 ± 0.023 (45.62)
TIR domain	20	672	231.9	41	21	0.0273 ± 0.0142	0.0255 ± 0.0100	0.151 ± 0.024 (37.30)

ND, not determined.
[a] Number of samples.
[b] Total length of sequenced region.
[c] Number of segregating sites.
[d] Number of segregating silent sites.
[e] Estimates based only on silent sites.
[f] Divergence to *A. lyrata* based on silent sites, with standard errors provided, and the number of silent differences given in parentheses.

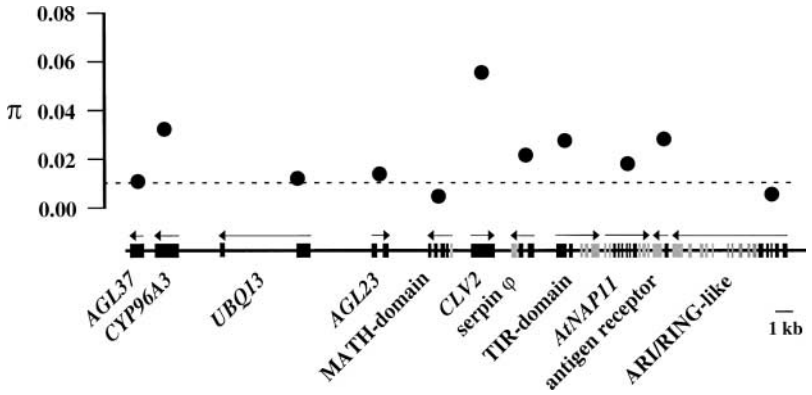

FIGURE 1.—Levels of silent site nucleotide diversity at *CLV2* and flanking loci. The dashed line shows the mean level of nucleotide diversity (π) in previously studied genes of *A. thaliana*. Sequenced and unsequenced exons are shown in thick solid or shaded bars, respectively. The line connecting *UBQ13* fragments spans the mitochondrial DNA insertion observed in Col-0. Arrows indicate each gene's orientation in the chromosome.

or rapid post-Pleistocene range expansion of this species (SHARBEL *et al.* 2000). However, only 2 genes—the putative serpin (Fu and Li $D = -1.6981$, $P < 0.05$) and the putative antigen receptor (Tajima's $D = -1.9246$, $P < 0.05$; Fu and Li $D^* = -2.2497$, $P < 0.01$)—show significantly negative values of at least one test statistic. In the latter case, this significant excess in low-frequency polymorphisms is largely due to the presence of a single divergent haplotype from the Moroccan Ita-0 ecotype.

In contrast, both *CLV2* and the TIR domain gene have consistently positive values of the Tajima and Fu and Li test statistics (Table 4A), but only the TIR domain gene was significantly positive (Fu and Li $D^* = +1.2984$, $P < 0.05$; $D = +1.6783$, $P < 0.01$). Loci with significant positive values of these test statistics have rarely been observed in previous studies of *A. thaliana*. Positive values of these test statistics are associated with an excess of intermediate-frequency polymorphisms. These data suggest that both of these genes may be evolving nonneutrally in a pattern consistent with balancing selection, but the power of these tests is limited at such small sample sizes.

Since our results indicated that the *CLV2* gene has the highest level of polymorphism among the 11 linked genes, we examined variation at this gene and its three closest neighbors in greater detail. We sequenced additional accessions at these loci to increase the number of sampled alleles to 19–21. The results from this expanded data set are consistent with the patterns observed with the smaller data set. The levels of nucleotide variation, the directions of the Tajima's *D* and Fu and Li's *D** and *D* tests statistics, and the results of the HKA tests against *Adh* are all comparable across the two data sets (Tables 3 and 4). The only difference is that with larger sample sizes, the value of Tajima's *D* is now significant for *CLV2* ($D = +1.752$, $P < 0.05$). This finding is consistent with previous analyses that indicated that augmenting sample sizes for sequenced alleles increases the power to detect significant deviations from the neutral-equilibrium model (SIMONSEN *et al.* 1995).

The positive value of Tajima's *D* in *CLV2* is associated with the presence of at least three distinct haplotype groups (I, II, and IV in Figure 3). These three haplogroups are found at moderate frequency, with the rarest haplogroup at ~15% frequency. Also, one haplotype (III in Figure 3) may have arisen from a recombination event between alleles belonging to groups II and IV. Alternatively, haplotype III, obtained from the Ita-0 accession, may represent an additional allelic class; this accession also bears more divergent alleles of several other loci in this region.

Intragenic and intergenic linkage disequilibrium at the *CLV2* region: Linkage disequilibrium, the nonrandom association of allelic polymorphisms, was surveyed for nucleotide polymorphisms both within and between genes in the *CLV2* region. The amount of linkage dis-

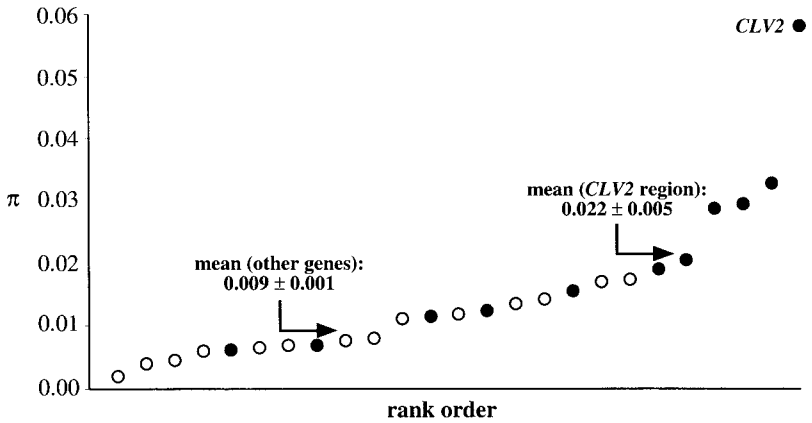

FIGURE 2.—Comparison of silent site nucleotide diversity between genes in the *CLV2* region and other *A. thaliana* genes. Previously studied genes (open circles) and genes at the *CLV2* region (solid circles) are ranked by π.

TABLE 4

Selection tests at genes in the *A. thaliana CLV2* genomic region

Gene	*n*	Tajima's *D*	Fu and Li *D**	Fu and Li *D*	HKA test[a]
A. Initial sampling at all loci					
AGL37	10	+0.3643	+0.0638	−0.0943	0.84
CYP96A3	9	−0.4734	−0.9824	−0.4615	0.08
UBQ13	10	−0.8460	−1.3910	−1.4250	0.56
AGL23	9	−0.7813	−1.0620	−1.2981	0.40
MATH domain	9	−0.8910	−0.6896	−1.1241	ND
CLAVATA2	12	+1.0678	+0.4589	+0.5229	0.01
serpin ϕ	10	−0.7173	−1.0888	−1.6981*	0.07
TIR domain	10	+0.1770	+1.2984*	+1.6783**	0.11
AtNAP11	10	−1.0890	−1.0947	−0.3432	0.03
Antigen receptor	10	−1.9246*	−2.2497**	−1.1371	0.04
ARI/RING-like	10	−0.4873	−0.1377	−0.3609	0.77
B. After additional sampling of CLV2 and the three nearest loci					
MATH domain	19	−1.0754	−0.7730	−1.1241	ND
CLAVATA2	21	+1.752*	+0.7833	+0.8744	0.02
serpin ϕ	20	−0.5189	−0.7007	−0.7085	0.10
TIR domain	20	+0.5945	+1.2805*	+1.4709*	0.16

ND, not determined. *$P < 0.05$; **$P < 0.01$.
[a] *P* values against reference neutral gene *Adh*.

equilibrium was estimated using the r^2 statistic (Hill and Robertson 1968) for nonsingleton sites, and the significance of pairwise disequilibrium comparisons was assessed with Fisher's exact test. In the smaller sampling of 9–12 accessions, 61% of intragenic comparisons are significant (a total of 1423 pairs of sites and a range of 6–379 per gene). The proportion of significant disequilibrium values for pairwise comparisons ranges from 10% for *AGL37* to 95% for *CYP96A3* and the TIR domain gene.

Larger sample sizes increase the power of detecting significant linkage disequilibrium, and this is demonstrated for four genes (the MATH domain gene, *CLV2*, the serpin pseudogene, and the TIR domain gene), which were examined in the expanded sample set of 19–21 ecotypes. The proportion of significant comparisons ranges from ~1 to ~30% of pairwise comparisons (Table 5). The levels of intragenic linkage disequilibrium can also be estimated using the Z_{nS} statistic (Kelly 1997). The levels of Z_{nS} are significantly higher than expected under neutrality for the *CLV2* gene ($P < 0.012$), the serpin pseudogene ($P < 0.04$), and the TIR domain gene ($P < 0.008$), as assessed by coalescent simulations that take into account the population recombination parameter estimated from the data.

The extent of disequilibrium between genes is evident in plots of r^2 as a function of physical distance. Across the entire 40-kb region, strong linkage disequilibrium ($r^2 = 1$) is observed even at distances of ~25 kb in the smaller data set (Figure 4A). Using data from the expanded sample set, strong levels of intergenic disequilibrium are also evident among *CLV2* and its nearest neighbors (Figure 4B). The distance plot shows strong linkage disequilibrium up to ~6 kb associated with correlations among *CLV2*, the serpin pseudogene, and the TIR domain locus.

Amino acid replacements at *CLV2*: In our sample of *CLV2* alleles, 22 of the 54 nucleotide polymorphisms code for amino acid replacements (Figure 3A); 20 of the substitutions occur in the LRRs, while two are in the cysteine-pair region preceding the LRRs (Figure 5). Proteins in the four allele classes differ by 7–15 amino acids. Although the majority of these replacements are fairly conservative, two to five of the differences between allele classes are due to radical substitutions (Figure 3B). The amino acid substitutions observed in our data set probably encompass much of the variation present within the species. Comparisons of the full-length *CLV2* sequence from the Col-0 (class I), Ws-0 (class II), and Ler-0 (class IV) ecotypes reveal only two additional amino acid replacements, one in the 18th LRR and one in the cysteine-pair region following the LRRs (Jeong *et al.* 1999).

DISCUSSION

Contrasting patterns of sequence variation across the *A. thaliana* CLV2 region: Molecular population genetic analyses of the *A. thaliana CLV2* region indicate that levels and patterns of nucleotide diversity can vary even among contiguous, closely linked genes. For example, although *CLV2* has the highest level of nucleotide variation in this region ($\pi = 0.0558$), the MATH domain gene has the lowest ($\pi = 0.0060$)—a nearly 10-fold reduction in diversity between adjacent genes. Similar patterns of differing nucleotide diversity levels among

A

```
                 1111111111112222222222233344445555555566888899
                 1230024556788990001445666701956770013699938789900
                 8606165912772812013926401399187403076380360427675 8
Col-0            CTCACCCGCAGGTACGTGAATGCGGCATTGACGATAAGAATCACCTCGCC
Lan-0            .....................................................
Fi-0             .....................................................
Cvi-0            A.TGT.TATT...T....G...T.ATGAA.......................
Bla-1            A.TGT.TATT...T....G...T.ATGAA.......................
Bs-1             A.TGT.TATT...T....G...T.ATGAA.......................
An-2             A.TGT.TATT...T.A..G...T.ATGAA.......................
Lisse2           T.TGT.TATT...T....G...T.ATGAA.......................
Sf-1             T.TGT.TATT...T....G...T.ATGAA.......................
Chi-1            ACTGT.TATT...T....G...T.ATGAA..............TAT.GA
Ws-0             ACTGT.TATT...T....G...T.ATGAA..............TAT.GA
Ita-0            T.TGT.TA.T...T....G.......G.AAGGA.AGCTC....TTC.....
Jl-3             ..TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Gr-3             ..TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
El-0             ..TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Bu-0             ..TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Kas-1            T.TGT.....AAGTTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Co-1             T.TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Lu-1             T.TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Pi-0             T.TGT.....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..
Ler-0            T.TGTG....AA.TTCCAGGCA.A.TGC..GGAGAGC..GCTG....T..

                 L S S  T  V   A V R D  V NF MQ    II K   MR   A
                 S G C  I  I   T M G N  I DI IR    VL T   VC   S
                               P          L
```

I II III IV

FIGURE 3.—Polymorphisms in the *CLV2* gene. (A) Table of nucleotide polymorphisms in 21 *A. thaliana* accessions. Positions of polymorphic sites are indicated at the top. All alleles are compared to the Col-0 reference sequence. Brackets denote the four allelic classes observed. For sites containing nonsynonymous substitutions, the amino acid polymorphisms are shown beneath the nucleotide polymorphisms; the first line shows the Col-0 residue, while subsequent lines show replacement residues. (B) Numbers of amino acid replacement polymorphisms within and among allele classes. Total numbers of replacements are above the diagonal. Replacements within an allele class are on the diagonal. Radical replacements are below the diagonal.

B

	I	II	III	IV
I	*0*	7	9	14
II	2	*2*	11	15
III	3	3	*0*	11
IV	4	4	5	*1*

linked genes have also been observed in a 400-kb region around the *FRI* gene (HAGENBLAD and NORDBORG 2002), in a 170-kb region around the *MAM1* gene (HAUBOLD *et al.* 2002), and in a 20-kb region around *RPS5* (TIAN *et al.* 2002). The variation in estimates of sequence polymorphism even among contiguous genes also suggests that surveys of nucleotide diversity may require more extensive sampling in a given genomic region to arrive at better estimates of region-specific polymorphism levels.

There also appear to be dramatic changes in the patterns of nucleotide variation observed among neighboring loci in the *CLV2* region. Both *CLV2* and the TIR domain locus, for example, have positive levels of Tajima's *D*, consistent with an excess of intermediate frequency polymorphisms in the sampled alleles. These two loci, however, are surrounded by and interspersed with genes that display negative levels of Tajima's *D*, indicating an excess of low-frequency polymorphisms for these linked loci. These results suggest that levels and patterns of variation are remarkably gene-specific even among closely linked *A. thaliana* nuclear genes.

Linkage disequilibrium levels appear to be extensive across the *CLV2* region. In this 40-kb region, disequilib-

TABLE 5

Intragenic linkage disequilibrium at *A. thaliana CLV2* and nearest genes

Gene	*n*	No. of sites[a]	No. of comparisons	Significant comparisons[b]	Z_{nS}
MATH domain	19	11	55	18	0.274 ($P < 0.110$)
CLAVATA2	21	40	780	409	0.455 ($P < 0.012$)
serpin ϕ	20	25	300	194	0.408 ($P < 0.040$)
TIR domain	20	38	703	666	0.839 ($P < 0.008$)

[a] Excluding singleton mutations.
[b] Significant at the $P < 0.05$ level using Fisher's exact test.

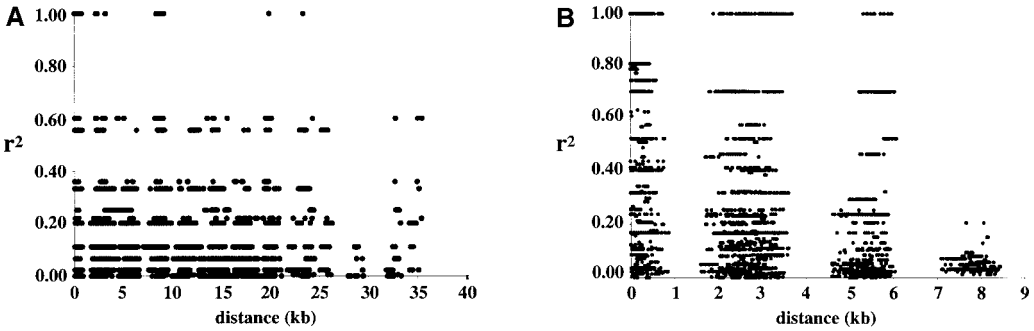

FIGURE 4.—Linkage disequilibrium in the *CLV2* genomic region. All site comparisons separated by <1 kb are intragenic; the remainder are intergenic. (A) Linkage disequilibrium across all 11 genes in the *CLV2* region determined from 8 accessions. (B) Linkage disequilibrium among the MATH domain, *CLV2*, serpin, and TIR domain genes determined from 19 accessions.

rium is observed both intra- and intergenically, and strong disequilibrium can extend to ~25 kb. There is also evidence for correlation of allele genealogies among some of the linked genes (K. A. SHEPARD, unpublished observations). This correlation in gene genealogies, however, is not observed between genes that are farther apart and can also disappear between adjacent loci. The *CLV2* gene and the MATH domain locus immediately upstream, for example, display weaker correlation in genealogies among the sampled alleles (K. A. SHEPARD, unpublished observations).

Several of the genes in the *CLV2* region appear to contain two or more distinct haplotype groups (see, for example, Figure S1 in Supplementary Information at http://www.genetics.org/supplemental/). The presence of two distinct allele groups, commonly referred to as allelic dimorphism, has been observed in previous studies of

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
	0000	11111	1111	111	2	2	22	222	3	3	3
	8888	23334	4566	778	0	2	44	688	0	2	7
	0789	52492	8728	284	7	7	47	445	3	6	4
I	**L**SQI	**S**	**T**RV**A**V	RDVN	NFV	M	Q	II	**K**LM	**R**	S	A
II	.G..	.	**I**....	..ID	.I.	.	.	..	**.**..	.	.	.
IIa	.G..	.	**I**..**T**.	..ID	.I.	.	.	..	**.**..	.	.	.
IIb	**S**G..	.	**I**....	..ID	.I.	.	.	..	**.**..	.	.	.
III	.G..	.	.....	...D	.L.	I	R	VL	**T**..	**C**	.	.
IV	.G..	.	..**I**P**M**	GN.D	.L.	.	R	VL	..V	.	.	S
IVa	.G..	**C**	..**I**P**M**	GN.D	.L.	.	R	VL	..V	.	.	S
Lyr	.GEL	.	.S.**G**.	K.IG	ILL	L	R	VL	.F.	.	R	.
		β	α	β	α	β		ββ	α			

FIGURE 5.—Predicted amino acid replacements encoded by *CLV2* alleles. Replacement changes based on predicted protein sequences for representative alleles within haplotype classes as designated in Figure 3. IIa is the An-2 allele. IIb corresponds to the Chi-1 and the Ws-0 alleles. Lyr is the *A. lyrata* ortholog. Radical amino acid substitutions are indicated in boldface type. The numbers for each LRR are shown on top in italics, and the amino acid positions for each replacement are also numbered, as designated by JEONG *et al.* (1999). Replacements in the β-strand/β-turn region as predicted by the conserved sequence motif xxLxLxx are designated as "β." "α" denotes replacements in possible α-helical regions of the Col-0 accession as predicted by SSpro2 (BALDI *et al.* 1999).

A. thaliana (KAWABE *et al.* 1997; PURUGGANAN and SUDDITH 1998). For most *A. thaliana* nuclear genes, allelic dimorphism appears to be readily accounted for by a model of neutral evolution with no recombination and may represent the remnants of ancestral population structure (KUITTINEN and AGUADÉ 2000; AGUADÉ 2001). In a few instances, however, the elevated nucleotide variation associated with these highly divergent alleles is more compatible with balancing selection at a locus (STAHL *et al.* 1999; OLSEN *et al.* 2002; TIAN *et al.* 2002).

The long-range decay of linkage disequilibrium is expected in *A. thaliana*, a predominantly selfing species with a reduced effective recombination rate. Unlike in *D. melanogaster* or *Z. mays*, where disequilibrium decays in scales of ~1 kb, linkage disequilibrium in *A. thaliana* can persist up to 250 kb (NORDBORG *et al.* 2002). Given reduced recombination in *A. thaliana*, balanced polymorphisms may be expected to display high levels of variation and maintenance of alternate haplotypes over longer genomic scales in *A. thaliana* (NORDBORG *et al.* 1996), comparable to the persistence of disequilibrium in this selfing species (NORDBORG *et al.* 2002).

Evidence for balancing selection in the *CLV2* genomic region: The reported high level of polymorphism at the *CLV2* meristem regulatory gene (JEONG *et al.* 1999) first suggested the possibility that alleles at this locus or a nearby linked gene may be maintained as a balanced polymorphism in *A. thaliana*. A survey of the levels and patterns of variation among 11 linked genes centered on *CLV2* was undertaken to dissect the evolutionary forces acting on this 40-kb genomic region. Three aspects of the levels and patterns of nucleotide diversity at *CLV2* are noteworthy. First, the level of silent site nucleotide diversity at this developmental gene is about fivefold higher than those of typical *A. thaliana* nuclear genes; this is one of the highest levels of variation thus far reported in this species. The level of variation at *CLV2* is also significantly higher than that of the reference neutral gene *Adh* (HKA test, $P < 0.01$). Second, Tajima's *D* is significantly positive for this gene ($P <$

0.01), which indicates an excess of intermediate-frequency polymorphisms. Third, the level of intragenic linkage disequilibrium at this locus is significantly higher than that predicted by a neutral-equilibrium model under limited recombination (Z_{nS} statistic, $P < 0.012$).

Three alternative scenarios may explain this pattern of diversity at the *CLV2* gene. One possibility is a duplication at this locus, which could explain the distinct haplogroups, high variation, and intragenic linkage disequilibrium. There is no evidence, however, for a recent duplication of *CLV2* or any of the genes flanking it in the Arabidopsis genome. Moreover, we find no evidence of duplication heterozygosity in different *A. thaliana* ecotypes (K. A. Shepard, unpublished observation). A second scenario is that contemporary or ancestral geographical subdivision can also result in the observed pattern. Detailed analysis of *A. thaliana* ecotypes using genome-wide markers, however, does not reveal any strong geographical subdivision within this species (Sharbel *et al.* 2000). Molecular population genetic analyses of various genes do reveal the sporadic presence of allelic dimorphism compatible with ancestral subdivision (Kawabe *et al.* 1997; Miyashita *et al.* 1998). The levels of nucleotide variation at these loci, however, do not show marked elevation, nor do they display significant positive levels of either Tajima's or Fu and Li's statistics. These observations suggest that diversity at these genes, but not at *CLV2*, is compatible with neutral evolution under no recombination (Aguadé 2001). The third alternative compatible with the observed levels and patterns of nucleotide variation at *CLV2* is that this gene harbors a balanced polymorphism. Similar patterns have been noted in other loci that unequivocally harbor balanced polymorphisms, including the *Rpm1* (Stahl *et al.* 1999) and *RPS5* (Tian *et al.* 2002) disease-resistance genes. It should be noted that the balanced polymorphism at *CLV2* may not be incompatible with the possibility of ancestral geographical subdivision. The *CLV2* haplogroups, for example, may have originated from locally adapted, geographically distinct ancestral populations (Charlesworth *et al.* 1997) and may be currently maintained by local selection on alternate alleles despite the widespread post-Pleistocene dispersal of this species.

The only other gene in this region that shows some evidence for balancing selection is the TIR domain gene located ~4 kb downstream of *CLV2*. This locus has significantly positive Fu and Li and Z_{nS} disequilibrium test statistics; unlike *CLV2*, however, this gene does not show significantly high intraspecific nucleotide variation compared to *Adh* (HKA test, $P < 0.7$). The pattern at the TIR domain gene may simply result from linkage with a balanced polymorphism at *CLV2*, as is suggested by the allele groups shared among these loci (see Figures 3 and S5 at http://www.genetics.org/supplemental/). Alternatively, balancing selection may be acting independently on the TIR domain gene. The sequence of this gene is similar to the TIR portion of the *RPS4* disease-resistance gene (Gassmann *et al.* 1999), but it lacks the nucleotide binding site and LRRs characteristic of proteins encoded by *RPS4* and other TIR-containing disease-resistance genes in plants. If balancing selection is acting directly on this gene, and not as a correlated effect from putative balanced polymorphisms at *CLV2*, it may be associated with as yet uncharacterized disease-resistance functions at this locus.

While levels of nucleotide variation are predicted to be highest immediately surrounding a balanced polymorphism, an elevated level of variation may also be expected in a more extended genomic region of a predominantly selfing species. This predicted pattern is also observed by the high level of nucleotide variation among the 11 linked genes in the *CLV2* genomic region. There is a twofold increase in estimates of variation between loci in the *CLV2* region and a set of 14 other *A. thaliana* genes. There is no accompanying increase in nucleotide divergence estimates for these genes between *A. thaliana* and *A. lyrata*, compared to previously studied loci. This suggests that the increase in intraspecific nucleotide variation in this region is not the result of an increase in the neutral mutation rate.

Our results, however, indicate that while a wide window of enhanced neutral variation surrounds the putative balanced polymorphisms in *CLV2*, significant effects of selection on levels and patterns of sequence diversity appear confined to genic scales. The localized nature of the effects of balanced polymorphisms in the predominantly selfing *A. thaliana* is paradoxical, although it has been observed at several loci. In the *RPS5* disease-resistance locus, significantly enhanced variation is observed surrounding the sequence junction that harbors the *RPS5* balanced indel polymorphism, but is not observed at adjacent loci within ~10 kb (Tian *et al.* 2002). Similarly, a balanced polymorphism at the *TFL1* inflorescence architecture gene is confined to the 1-kb promoter region, and increased diversity is not observed in either the *TFL1* coding region or the upstream *rps28* gene (Olsen *et al.* 2002). Finally, a replacement polymorphism associated with a Fast/Slow allozyme polymorphism at the *PgiC* locus is intragenically localized, spanning a region of only five exons and intervening introns (Kawabe *et al.* 2000). These results are consistent with our observations in the *CLV2* region that significant retained effects of balancing selection on levels and patterns of sequence diversity may be focused at specific genes and not at nearby linked loci.

The *CLV2* gene, and to some extent the TIR domain locus, are the only two genes that display departures from neutral-equilibrium predictions by several criteria: (i) significantly elevated levels of nucleotide variation, (ii) intermediate-frequency polymorphisms, and (iii) intragenic linkage disequlibrium. The other genes in the *CLV2* region may also have been affected by selection

at or near these loci, but do not retain consistent signatures of balancing or positive selective forces. This may reflect, in part, the relatively low power of some of the tests for selection (Simonsen *et al.* 1995). Loci may, for example, harbor balanced polymorphisms but the frequency of allele classes are not sufficiently high to provide a significant positive value of Tajima's *D*.

Functional consequences of the putative balanced polymorphism at *CLV2*: The functional consequences of natural allelic differentiation at *CLV2* remain unclear. The putatively balanced alleles at *CLV2* are associated with a large number of replacement polymorphisms, with 7–15 amino acid changes differentiating different allele groups. The distribution of amino acid replacements within LRRs suggests that some of these substitutions could affect the function of the CLV2 protein. Extracellular plant LRRs are characterized by the consensus amino acid sequence LxxL{xxLxLxx}NxLx GxI-PxxLGx, where L may also be isoleucine, valine, or phenylalanine. Plant-specific LRRs have not yet been crystallized; however, structural analyses of nonplant proteins predict that each LRR consists of a β-strand and an α-helix joined by loops. The alternating β-strands and α-helices yield a horseshoe-shaped structure in which parallel β-strands form a binding pocket for protein-protein interactions. The xxLxLxx motif forms a β-strand/β-turn with buried leucine residues and solvent-exposed variable residues (Kobe and Kajava 2001).

In our *CLV2* data set, three amino acid substitutions occur in the solvent-exposed residues in the β-strand/β-turn (Figure 5). Two of these mutations ($Thr_{125} \leftrightarrow$ Ile and $Arg_{148} \leftrightarrow$ Gly) are radical substitutions, while the third ($Ile_{244} \leftrightarrow$ Val) is conservative. Recent studies of cytoplasmic LRR proteins that confer disease resistance in plants have highlighted the functional importance of variation in solvent-exposed LRR residues. Evidence for diversifying selection on these residues has been observed in comparisons of paralogous disease-resistance genes within several species (Parniske *et al.* 1997; McDowell *et al.* 1998; Meyers *et al.* 1998; Wang *et al.* 1998; Bittner-Eddy *et al.* 2000; Dodds *et al.* 2001). In the CLV2 LRRs, we did not find support for diversifying selection as measured by K_a/K_s, the ratio of nonsynonymous to synonymous nucleotide substitution rates (data not shown). However, analysis of the P2 and p-B genes of flax indicates that less dramatic variation can also alter protein function. The predicted P2 and p-B proteins, which confer recognition of different rust strains, differ by only six solvent-exposed residues (Dodds *et al.* 2001). These results suggest that the variation we observe in the CLV2 β-strand/β-turn might have functional consequences.

The majority of amino acid replacements in CLV2 are located in the interstrand regions of the LRRs. Of these 14 replacements, only 2 are predicted to reside in helical motifs, suggesting that the remainder are found in loops (Figure 5). Although the structure-function relationships in the interstrand regions are less understood, residues in loop regions can clearly affect LRR protein function. Studies of natural variation at *RPS2*, an *A. thaliana* disease-resistance gene, have shown that six amino acid differences between the Col-0 (resistant) and Po-1 (susceptible) alleles are sufficient to alter pathogen recognition (Banerjee *et al.* 2001). Two of these mutations are found in both resistant and susceptible alleles in other ecotypes (Caicedo *et al.* 1999), suggesting that they are not specificity determinants. The remaining four mutations, which are located in interstrand regions of the RPS2 protein, indicate that residues outside the β-strand/β-turn can lead to functional diversification among alleles. This result is not surprising, as structural analyses have shown that ligand binding to other types of LRR proteins often involves contacts in the loops as well as in the β-strands (reviewed by Kobe and Kajava 2001).

If, indeed, some of these replacement substitutions are maintained as balanced polymorphisms, the mechanism of selection is puzzling in light of what little is known about *CLV2*'s role in plant development. Although there is compelling genetic evidence that the proteins encoded by the three *CLAVATA* genes act together to regulate shoot meristem growth, the exact constituents of and binding relationships among the receptor and ligand multimers are unclear. Of the three characterized *CLAVATA* genes, *clv2* mutant alleles show the weakest shoot meristem phenotypes (Kayes and Clark 1998). The mild *clv2* phenotype may indicate that this gene is not a crucial regulator of meristem function in Arabidopsis; however, mutations in the *fasciated ear2* gene, a putative *CLV2* ortholog, have dramatic effects on maize inflorescence morphology (Taguchi-Shiobara *et al.* 2001). Moreover, unlike the meristem-specific phenotypes of *clv1* and *clv3* mutants, *clv2* plants show pleiotropic effects on pedicel, stamen, and gynoecium development (Kayes and Clark 1998). Finally, in contrast to the narrow, meristematic expression domains of *CLV1* and *CLV3*, the broad expression pattern of *CLV2* in the shoot (Jeong *et al.* 1999) suggests that CLV2 may interact with additional proteins in other parts of the plant.

We therefore propose two hypotheses that might explain the putative balancing selection on the *CLV2* locus. First, CLV2 might act as a modulator of shoot meristem growth, with different alleles enhancing or reducing the strength of signaling through the CLAVATA complex. This modulation might be accomplished by variation in the accumulation of CLV1 protein in the plasma membrane or by alterations in the affinity of the complex for the multimeric CLV3 ligand. Such modulation could have direct effects on fitness-related traits such as flower number. Alternatively, balancing selection may act on pleiotropic functions of CLV2 that involve currently unidentified binding partners. Characterizing phenotypic, ecologically relevant variation associated

with alleles at *CLV2* will strengthen the argument of balancing selection at this locus.

The authors thank Brandon Gaut, Ken Olsen, Mark Ungerer, Montserrat Aguadé, two anonymous reviewers, and members of the Purugganan laboratory for helpful comments, Outi Savolainen and Helmi Kuittinen for providing *A. lyrata* seed, and Juergen Kroymann for providing preprints of relevant manuscripts. The authors are also grateful to the NCSU Phytotron for providing growth facilities and the NCSU Genome Research Laboratory for sequencing facilities. This work was funded by a grant from the National Science Foundation Integrated Research Challenges in Environmental Biology program to M.D.P., J. Schmitt, and T.F.C. Mackay, and an Alfred P. Sloan Foundation Young Investigator Award to M.D.P.

LITERATURE CITED

Abbot, R. J., and M. F. Gomez, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62:** 411–418.

Aguadé, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. Mol. Biol. Evol. **18:** 1–9.

Baldi, P., S. Brunak, P. Frasconi, G. Soda and G. Pollastri, 1999 Exploiting the past and the future in protein secondary structure prediction. Bioinformatics **15:** 937–946.

Banerjee, D., X.-C. Zhang and A. F. Bent, 2001 The leucine-rich repeat domain can determine effective interaction between *RPS2* and other host factors in Arabidopsis *RPS2*-mediated disease resistance. Genetics **158:** 439–450.

Bittner-Eddy, P. D., I. R. Crute, E. B. Holub and J. L. Beynon, 2000 *RPP13* is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. Plant J. **21:** 177–188.

Brand, U., J. C. Fletcher, M. Hobe, E. M. Meyerowitz and R. Simon, 2000 Dependence of stem cell fate in *Arabidopsis* on a feedback loop regulated by CLV3 activity. Science **289:** 617–619.

Caicedo, A. L., B. A. Schaal and B. N. Kunkel, 1999 Diversity and molecular evolution of the *RPS2* resistance gene in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **96:** 302–306.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided population. Genet. Res. **70:** 155–174.

Dodds, P. N., G. J. Lawrence and J. G. Ellis, 2001 Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the *P* and *P2* rust resistance specificities in flax. Plant Cell **13:** 163–178.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. Genome Res. **8:** 175–185.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gassmann, W., M. E. Hinsch and B. J. Staskawicz, 1999 The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J. **20:** 265–277.

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

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

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Innan, H., and W. Stephan, 2000 The coalescent in an exponentially growing metapopulation and its application to *Arabidopsis thaliana*. Genetics **155:** 2015–2019.

Innan, H., F. Tajima, R. Terauchi and N. T. Miyashita, 1996 Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics **143:** 1761–1770.

Jeong, S., A. E. Trotochaud and S. E. Clark, 1999 The Arabidopsis *CLAVATA2* gene encodes a receptor-like protein required for the stability of the CLAVATA1 receptor-like kinase. Plant Cell **11:** 1925–1933.

Kawabe, A., and N. T. Miyashita, 1999 DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. Genetics **153:** 1445–1453.

Kawabe, A., H. Innan, R. Terauchi and N. T. Miyashita, 1997 Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. Mol. Biol. Evol. **14:** 1303–1315.

Kawabe, A., K. Yamane and N. T. Miyashita, 2000 DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics **156:** 1339–1347.

Kayes, J. M., and S. E. Clark, 1998 *CLAVATA2*, a regulator of meristem and organ development in *Arabidopsis*. Development **125:** 3843–3851.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Kobe, B., and A. V. Kajava, 2001 The leucine-rich repeat as a protein recognition motif. Curr. Opin. Struct. Biol. **11:** 725–732.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-Dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Kuittinen, H., and M. Aguadé, 2000 Nucleotide variation at the *CHALCONE ISOMERASE* locus in *Arabidopsis thaliana*. Genetics **155:** 863–872.

Kumar, S., K. Tamura, I. B. Jakobsen and M. Nei, 2001 MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:** 1244–1245.

Long, A. D., R. F. Lyman, C. H. Langley and T. F. C. Mackay, 1998 Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. Genetics **149:** 999–1017.

McDowell, J. M., M. Dhandaydham, T. A. Long, M. G. M. Aarts, S. Goff *et al.*, 1998 Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of Arabidopsis. Plant Cell **10:** 1861–1874.

Meyers, B. C., K. A. Shen, P. Rohani, B. S. Gaut and R. W. Michelmore, 1998 Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell **11:** 1833–1846.

Miyashita, N. T., 2001 DNA variation in the 5′ upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. Mol. Biol. Evol. **18:** 164–171.

Miyashita, N. T., A. Kawabe, H. Innan and R. Terauchi, 1998 Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. Mol. Biol. Evol. **15:** 1420–1429.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser. B **263:** 1033–1039.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory *et al.*, 2002 The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. **30:** 190–193.

Olsen, K. M., A. Womack, A. R. Garrett, J. I. Suddith and M. D. Purugganan, 2002 Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics **160:** 1641–1650.

Parniske, M., K. E. Hammond-Kosack, C. Golstein, C. M. Thomas, D. A. Jones *et al.*, 1997 Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. Cell **91:** 821–832.

Purugganan, M. D., and J. I. Suddith, 1998 Molecular population genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. Proc. Natl. Acad. Sci. USA **95:** 8130–8134.

Purugganan, M. D., and J. I. Suddith, 1999 Molecular population genetics of floral homeotic loci: departures from the equilibrium-

neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics **151:** 839–848.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Richman, A., 2000 Evolution of balanced genetic polymorphisms. Mol. Ecol. **9**: 1953–1963.

Rivas, S., T. Mucyn, H. A. van den Burg, J. Vervoort and J. D. G. Jones, 2002 An ~400 kDa membrane-associated complex that contains one molecule of the resistance protein Cf-4. Plant J. **29:** 783–796.

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

Rozen, S., and H. J. Skaletsky, 1998 Primer3 (http://www-genome.wi.mit.edu/genome_software/other/primer3.html).

Sharbel, T. F., B. Haubold and T. Mitchell-olds, 2000 Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol. Ecol. **9:** 2109–2118.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman and J. Bergelson, 1999 Dynamics of disease resistance polymorphism at the *Rpm1* locus of Arabidopsis. Nature **400:** 667–671.

Sun, C.-W., and J. Callis, 1993 Recent stable insertion of mitochondrial DNA into an Arabidopsis polyubiquitin gene by nonhomologous recombination. Plant Cell **5:** 97–107.

Taguchi-Shiobara, F., Z. Yuan, S. Hake and D. Jackson, 2001 The *fasciated ear2* gene encodes a leucine-rich repeat receptor-like protein that regulates shoot meristem proliferation in maize. Genes Dev. **15:** 2755–2766.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Tian, D., H. Araki, E. Stahl, J. Bergelson and M. Kreitman, 2002 Signature of balancing selection in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **99:** 11525–11530.

Trotochaud, A. E., T. Hao, G. Wu, Z. Yang and S. E. Clark, 1999 The CLAVATA1 receptor-like kinase requires CLAVATA3 for its assembly into a signaling complex that includes KAPP and a Rho-related protein. Plant Cell **11:** 393–405.

Trotochaud, A. E., S. Jeong and S. E. Clark, 2000 CLAVATA3, a multimeric ligand for the CLAVATA1 receptor-kinase. Science **289:** 613–617.

Uyenoyama, M. K., 2000 Mutational origin of new mating type specificities in flowering plants. Genes Genet. Syst. **75:** 305–311.

Wang, G.-L., D.-L. Ruan, W.-Y. Song, S. Sideris, L. Chen *et al.*, 1998 *Xa21D* encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **10:** 765–779.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: M. Aguadé