Letter to the Editor

The Coalescent in an Exponentially Growing Metapopulation and Its Application to *Arabidopsis thaliana*

Hideki Innan and Wolfgang Stephan

Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received December 10, 1999
Accepted for publication April 3, 2000

NATURAL selection affects the amount and pattern of DNA variation maintained in a population. However, it is very difficult to find evidence for natural selection from data of nucleotide sequence variation because there are many other factors affecting DNA polymorphism. One of the main factors is the demographic history of a population, including changes in population size and population structure.

To infer natural selection from DNA polymorphism data, Tajima's *D* test (Tajima 1989) is often used. Tajima's *D* statistic is based on the difference between two estimates of the amount of variation. One estimate is obtained from the number of segregating sites (Watterson 1975) and the other is based on the average number of pairwise differences (Nei and Li 1979; Tajima 1983). In a constant-size neutral equilibrium population, the expectation of Tajima's *D* is nearly zero because the expectations of both estimates are the same. When some kind of balancing selection is acting, Tajima's *D* tends to be positive. On the other hand, purifying selection can generate negative values of Tajima's *D*.

However, as discussed in Tajima (1993) and Charlesworth *et al.* (1993), changes of population size can also affect Tajima's *D*. In a population with decreasing size, the expectation of Tajima's *D* is positive, while a negative Tajima's *D* is predicted for a population with increasing size. Therefore, when the observed value of Tajima's *D* deviates significantly from 0, it is very difficult to know what the cause of such a deviation is. To understand the mechanism maintaining nucleotide variation in a population, it is important to distinguish these two factors, natural selection and changes of population size.

In this note, the effects of these two factors are successfully distinguished in the wild plant, *Arabidopsis thaliana*. *A. thaliana* is a highly selfing species with $<1\%$ outcrossing (Abbott and Gomes 1989). It is known that the *A. thaliana* natural population consists of a number of colonies. Sampled individuals from a colony are called ecotypes. There is almost no genetic variation within each colony (Todokoro *et al.* 1995; Bergelson *et al.* 1998), suggesting that the members of each colony may share a very recent common ancestor and that migration among colonies is rare. Thus, intraspecific variation in *A. thaliana* depends largely on variation among ecotypes. Usually, population studies of this species are conducted for a sample of ecotypes.

It is believed that the *A. thaliana* population expanded recently, originating from an area of the Himalayas (Price *et al.* 1994). Innan *et al.* (1996) investigated DNA sequence polymorphism of the *Adh* region and supported this idea because a number of rare variants with a negative Tajima's *D* value were observed. A lack of association between the genealogy and geographic origin of the sampled ecotypes is also consistent with a recent expansion of the *A. thaliana* population. Similar results were obtained for other nuclear gene regions (*e.g.*, *ChiA*, Kawabe *et al.* 1997; and *CAL*, Purugganan and Suddith 1998). At this time, data for eight gene regions are available, of which six loci exhibit a negative Tajima's *D* value. If we exclude *Rpm*1, where strong overdominant selection may be acting (Stahl *et al.* 1999), the average of Tajima's *D* is -1.34 (SD $= 0.90$), supporting the idea of a recent rapid expansion of *A. thaliana*.

To further investigate the demographic history of *A. thaliana*, we introduce a coalescent model of a metapopulation with *N* colonies. We assume that the size of all colonies is the same and that there is no migration among colonies. We also assume that a colony appears and gets extinct at fixed rates and that the number of colonies is constant. Consider a case where *n* ecotypes are sampled. To study the genealogical relationship among the ecotypes, we apply the theory of coalescence processes (Griffiths 1980; Kingman 1982; Hudson 1983; Tajima 1983) in the following way. For a coalescent among *n* genes, let t_i ($i = 1, 2, 3, \ldots, n - 1$) be the time during which $i + 1$ genes coalesce into i genes. Then, the distribution of t_i is given by

Corresponding author: Hideki Innan, Department of Biology, University of Rochester, Rochester, NY 14627.
E-mail: hi_innan@hotmail.com

$$F(t_i) = \frac{i(i+1)}{2Ng} e^{-[i(i+1)/2Ng]t_i}, \qquad (1)$$

where g is a scaling factor. g is defined such that Ng is the average time until two gene copies from two different, randomly sampled colonies coalesce. Clearly, g is affected by the rates of colonization and extinction. In *A. thaliana*, g is unknown.

Next, we consider a metapopulation whose number of colonies is increasing with a constant rate r per g generations. Let N_0 be the present number of colonies. Following Slatkin and Hudson (1991), the coalescent time (t_i) is given by

$$t_i = \ln\left[1 + N_0 r e^{-\tau_i} \frac{-2}{i(i+1)} \ln(U)\right], \qquad (2)$$

where

$$\tau_i = \sum_{k=i+1}^{n} t_k$$

and U is a random variable with a uniform distribution in the interval (0, 1) (Slatkin and Hudson 1991). The unit of t_i is $1/r$ generations. The coalescent process among the sampled ecotypes is described by a single parameter, N_0r. Using these results and assuming that the *A. thaliana* population has expanded exponentially, we can construct random genealogies among *A. thaliana* ecotypes in an exponentially growing metapopulation.

To estimate the demographic parameter N_0r, we analyzed the amplified fragment length polymorphism (AFLP) data of Miyashita *et al.* (1999). It is reasonable to use AFLP data for this purpose because the anonymous fragments detected in this analysis presumably cover the whole genome. In Miyashita *et al.* (1999), a total of 472 distinct fragments were observed for 38 ecotypes sampled from around the world. Their observation of a lack of linkage disequilibrium among the fragments suggests that almost all fragments follow independent genealogical histories.

The AFLP data by Miyashita *et al.* (1999) consist of 38 ecotypes, including an apparent outlier, ecotype Fl-3. For Fl-3, the observed number of fragments is about twice that of the remaining 37 ecotypes. Fl-3 appears to be an allotetraploid hybrid between *A. thaliana* and one of its close relatives (N. T. Miyashita and A. Kawabe, personal communication). In the following analysis, Fl-3 is therefore excluded. For the remaining 37 ecotypes, the average proportion of shared fragments, $\tilde{\tilde{F}}$, was 0.839, and nucleotide diversity (π) was estimated to be 0.010. It is important that the observed frequency spectrum looks much smoother (Figure 2 in Miyashita *et al.* 1999) than that of some particular gene regions (for instance, *Adh* and *ChiA*; Innan *et al.* 1996; Kawabe *et al.* 1997), supporting the idea that almost all observed fragments are independent.

Assuming that the number of *A. thaliana* colonies has increased exponentially, we can find an estimate of the demographic parameter, N_0r, by computer simulation using the least-squares method. The procedure of simulating AFLPs in a constant-size population was developed by Innan *et al.* (1999) and Miyashita *et al.* (1999). The simulation method for an exponentially growing population is similar to that of a constant-size population. First, a random ancestral sequence (genome) with the length M bp is constructed. This ancestral sequence consists of four nucleotides, A, T, G, and C, at equal frequencies. Innan *et al.* (1999) demonstrated that the effect of GC content on the AFLP polymorphism is very small if intraspecific variation is considered. Next, the sequence is divided into m parts with the length M/m. This means that the whole genome is assumed to have m independent genealogies. Recombination rate, genome size, and the number of chromosomes are related to m (see below). m random genealogies for n sequences are constructed and branch lengths are determined by Equation 2. Following a given genealogy for each part, random mutations are placed on the ancestral sequence. The number of mutations on a branch follows a Poisson distribution with mean $\mu TM/m$, where μ is the mutation rate per site per g generations and T is the length of the branch measured in generations. Then, we have n descendant sequences, from which we obtain the AFLP fragments according to the primer sequences. Finally, the lengths of the detected fragments are scored for n descendants.

In our simulations, $n = 37$ is assumed. The mutational parameter, $\theta_0 = 4N_0\mu$, is adjusted to produce the average proportion of pairwise shared bands, $\tilde{\tilde{F}} = 0.839$. M is assumed to be 1000 kb. This length is chosen to generate ~23 fragments per haploid individual per primer pair (as observed by Miyashita *et al.* 1999). This requires that the number of selective bases of the two primers is one. Since the number of chromosomes in a haploid genome of *A. thaliana* is five, we chose $m = 5$. It should be noted that m does not affect the expected AFLP distribution. In the *A. thaliana* natural population, m may be >5 because of recombination.

Simulating 2000 replicates for a given value of N_0r, we obtained the expected frequency spectrum of AFLPs. N_0r was varied in the range of 0–1.5 with an interval of 0.1, and the expected distribution of the frequency spectrum was obtained for each N_0r. Using these expectations, the fit to the observed spectrum (Miyashita *et al.* 1999) was investigated by the least-squares method. To estimate N_0r, we fitted a fourth-order polynomial to the discrete data points. Using this procedure, the minimum value of the sum of squares was obtained for $N_0r = 0.57$. We also used χ^2 to measure the goodness of fit instead of the sum of squares. In this case, the best fit was also found for $N_0r = 0.57$. Figure 1 shows both the observed spectrum of *A. thaliana* AFLPs and the expected one with $N_0r = 0.57$. The observed frequency spectrum is in excellent agreement with the expectation.

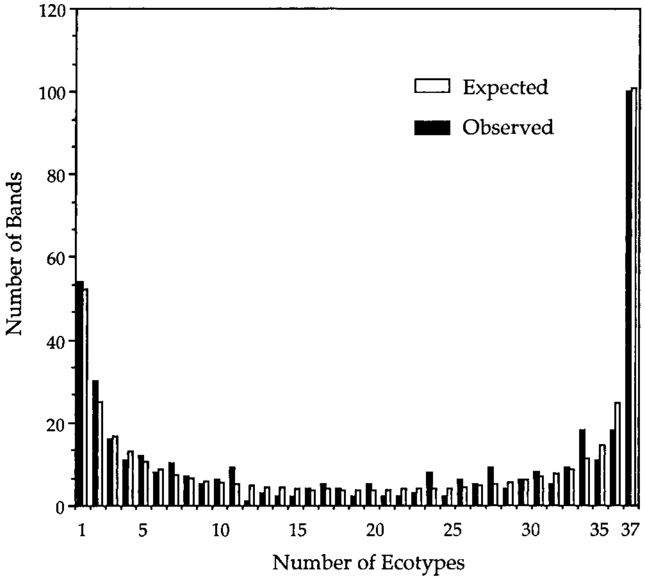

Figure 1.—The AFLP frequency spectrum in *A. thaliana*. Solid boxes show the observed spectrum and open boxes the expected one in an exponentially growing population with $N_0r = 0.57$. N_0r was estimated by fitting a fourth-order polynomial to the discrete data points. The minimum point of the curve was obtained for $N_0r = 0.57$.

Using this estimate of $N_0r = 0.57$, we also investigated the properties of Tajima's *D* (Tajima 1989) and Fu and Li's *D** tests (Fu and Li 1993) in an exponentially growing population. In our simulations, neutral genealogies without recombination (Hudson 1983, 1990; Tajima 1983, 1989) were used. Figure 2A shows the distribution of Tajima's *D* for $n = 17$. The filled squares represent the distribution of Tajima's *D* under a constant-population-size model ($N_0r = 0$) and the open squares represent that for an exponentially growing population with $N_0r = 0.57$. It is known that the distribution of Tajima's *D* under a constant-population-size model approximately follows the beta distribution with mean = 0 and variance = 1 (Tajima 1989). As shown in Figure 2A, the distribution with $N_0r = 0.57$ is skewed toward negative values. The mean is −0.3 and its variance is 0.55, which is smaller than that under a constant-population-size model. Accordingly, it is expected that the confidence limits of the distribution may be very different between these two models. Although the confidence limits for the negative values of Tajima's *D* are not much different from those of a constant-size model, those for the positive values are much smaller than those under a constant-size model when $N_0r = 0.57$. This implies that Tajima's *D* does not follow a beta distribution in growing populations. Similar results were obtained for Fu and Li's *D** test (Figure 2B). The peak of the distribution for $N_0r = 0.57$ moves toward negative values and the distribution is very different from that in a constant-size model especially for positive values.

The *P* values for Tajima's *D* and Fu and Li's *D** tests were reexamined for eight nuclear loci in *A. thaliana*, assuming $N_0r = 0.57$ (Table 1). As expected from Figure 2A, when Tajima's *D* is <-1.5, the *P* values with $N_0r = 0.57$ are not much different from those of a constant-size

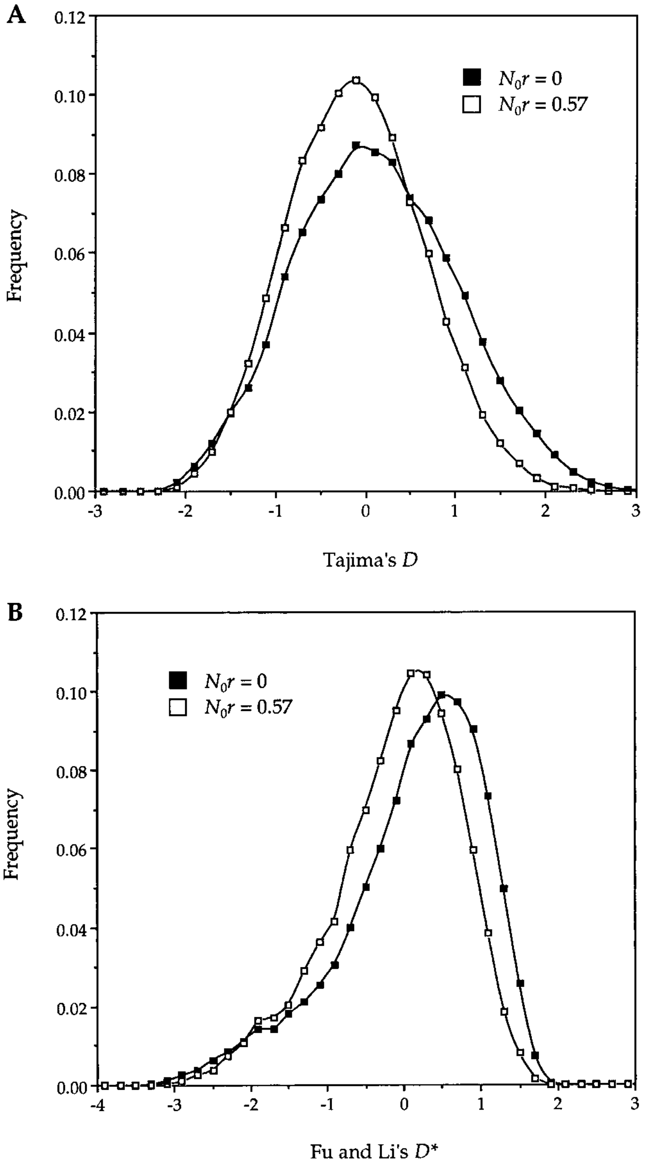

Figure 2.—The distributions of Tajima's *D* and Fu and Li's *D** statistics. They were obtained by computer simulation, in which $n = 17$ was assumed. Since in half of the genes in Table 1, the sample size is 17, simulation results for $n = 17$ are shown. In the simulation for a constant-size population, $4N_0\mu$ is assumed to be 20 so that the expectation of the average number of pairwise differences is 20, which appears to be a typical value (see Table 1). In an exponentially growing population with $N_0r = 0.57$, $4N_0\mu = 28.6$ is assumed. The expectation of the average number of pairwise differences is also 20. (A) The solid squares show the distribution of Tajima's *D* in a constant-population-size model, and the open squares show the distribution in an exponentially growing population with $N_0r = 0.57$. (B) The distribution of Fu and Li's *D** in a constant-population-size model and that in an exponentially growing population with $N_0r = 0.57$.

TABLE 1

Summary of Tajima's *D* and Fu and Li's *D tests**

					Tajima's D			Fu and Li's D^*		
Locus	n	S	S_u	K	D	$P(N_0r = 0)$	$P(N_0r = 0.57)$	D^*	$P(N_0r = 0)$	$P(N_0r = 0.57)$
Adh	17	75	38	19.11	−0.58	0.606	0.734	−1.05	0.307	0.367
*AP*3	19	78	69	10.53	−2.18	0.007**	0.003**	−3.37	0.003**	0.001**
CAL	17	91	65	16.29	−1.68	0.061	0.052	−2.20	0.065	0.049*
ChiA	17	123	104	17.99	−2.16	0.006**	0.003**	−2.95	0.009**	0.003**
ChiB	17	82	22	19.62	−0.81	0.438	0.522	0.25	0.837	0.552
PI	16	67	58	10.66	−2.02	0.017*	0.009**	−2.88	0.009**	0.003**
*Rpm*1	27	88	6	40.90	3.07	<0.001**	<0.001**	1.39	0.011*	<0.001**
*Rps*2	9	36	7	13.44	0.08	0.867	0.716	0.86	0.322	0.156

n, number of sequences; S, number of segregating sites; S_u, number of unique segregating sites; and K, average number of pairwise differences. Data for *Adh* are from Innan *et al.* (1996); *AP*3 and *PI* from Purugganan and Suddith (1999); *CAL* from Purugganan and Suddith (1998); *ChiA* from Kawabe *et al.* (1997); *ChiB* from Kawabe and Miyashita (1999); *Rpm*1 from Stahl *et al.* (1999); and *Rps*2 from Caicedo *et al.* (1999). *Significant at 5% level; **significant at 1% level.

model. For positive Tajima's D, the P value decreases for $N_0r = 0.57$, suggesting that the power of Tajima's D with $N_0r = 0.57$ is stronger than that of a constant-size model. Similarly, when Fu and Li's D^* is <-2.0, the P values with $N_0r = 0.57$ are not different from those of a constant-size model. If Fu and Li's D^* is positive, the P value with $N_0r = 0.57$ is smaller than that of a constant-size model. In the case of Fu and Li's D^* test for the *CAL* locus, the neutral hypothesis is rejected at the 5% level only when a growing population with $N_0r = 0.57$ is assumed. For two cases, *i.e.*, Tajima's D for *PI* and Fu and Li's D^* for *Rpm*1, the P value becomes <1% when N_0r is changed from 0 to 0.57.

In studies of particular gene regions, an excess of rare polymorphic sites was detected, resulting in significantly large negative values of Tajima's D and Fu and Li's D^* (Kawabe *et al.* 1997; Purugganan and Suddith 1999). As possible causes, an expansion of the *A. thaliana* population and/or purifying selection against deleterious mutations were proposed. In this study, two neutrality tests, Tajima's D and Fu and Li's D^*, were conducted without the effect of population expansion. It was shown that the negative values of both test statistics for *AP3*, *ChiA*, and *PI* were still significant (Table 1). Furthermore, Fu and Li's D^* for *CAL* is significantly negative at the 5% level when $N_0r = 0.57$, although it is not significant in a constant-size model. This result suggests that purifying selection against deleterious mutations is acting in these regions. This is consistent with the observation of a relatively high level of singleton replacement polymorphism in *A. thaliana* (Kawabe *et al.* 1997; Purugganan and Suddith 1998, 1999).

In this study, we inferred the rate of population size expansion of *A. thaliana* from the frequency spectrum of AFLPs obtained by Miyashita *et al.* (1999). For $N_0r = 0.57$, the observed spectrum is in excellent agreement with the expected one (Figure 1). The reason for this may be that almost all fragments detected as AFLPs are independent of each other, which is consistent with the lack of linkage disequilibrium observed between microsatellite loci (Innan *et al.* 1997) and AFLP loci (Miyashita *et al.* 1999). It may suggest that population surveys using AFLP analysis are an appropriate method to understand the demographic history of a population. Furthermore, using this estimate of population growth, we reevaluated the critical values of two of the most important neutrality tests, Tajima's D and Fu and Li's D^*. Discrepancies between constant-size and exponentially growing populations were found for positive values of both test statistics.

We thank two reviewers, M. Aguadé and D. Carlini, for valuable comments. A. Kawabe, N. T. Miyashita, and E. A. Stahl provided data sets. This work was supported in part by National Science Foundation grant DEB-989679 and by funds from the University of Rochester.

LITERATURE CITED

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

Bergelson, J., E. Stahl, S. Dudek and M. Kreitman, 1998 Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics **148:** 1311–1323.

Caicedo, A. L., B. A. Schaal and B. N. Kunkel, 1999 Diversity and molecular evolution of the PRS2 resistance gene in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **96:** 302–306.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

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

Griffiths, R. C., 1980 Lines of descent in the diffusion approximation of neutral Wright-Fisher models. Theor. Popul. Biol. **17:** 37–50.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

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.

Innan, H., R. Terauchi and N. T. Miyashita, 1997 Microsatellite polymorphism in natural populations of the wild plant *Arabidopsis thaliana*. Genetics **146:** 1441–1452.

Innan, H., R. Terauchi, G. Kahl and F. Tajima, 1999 A method for estimating nucleotide diversity from AFLP data. Genetics **151:** 1157–1164.

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.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

Miyashita, N. T., A. Kawabe and H. Innan, 1999 DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. Genetics **152:** 1723–1731.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Price, R. A., J. D. Palmer and I. A. Al-Shehbaz, 1994 Systematic relationships of *Arabidopsis*: a molecular and morphological perspective, pp. 7–19 in *Arabidopsis*, edited by E. M. Meyerowitz and C. R. Somerville. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

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.

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. Genetics **123:** 585–595.

Tajima, F., 1993 Statistical analysis of DNA polymorphism. Jpn. J. Genet. **68:** 567–595.

Todokoro, S., R. Terauchi and S. Kawano, 1995 Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Jpn. J. Genet. **70:** 543–554.

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

Communicating editor: M. Aguadé