

## Molecular Evolution and Positive Selection of the Symbiotic Gene *NORK* in *Medicago truncatula*

Stéphane De Mita,<sup>1,2</sup> Sylvain Santoni,<sup>1</sup> Isabelle Hochu,<sup>1</sup> Joëlle Ronfort,<sup>1</sup> Thomas Bataillon<sup>1,3</sup>

<sup>1</sup> UMR 1097, Diversité et Génome des Plantes Cultivées, Montpellier, France

<sup>2</sup> UMR 5171, Génome, Populations, Interactions, Adaptation, Université Montpellier 2, Montpellier, France

<sup>3</sup> Bioinformatics Research Center, Department of Genetics and Ecology, University of Aarhus, Aarhus, Denmark

Received: 1 October 2005 / Accepted: 21 December 2005 [Reviewing Editor: Dr. Deborah Charlesworth]

**Abstract.** Understanding the selective constraints of partner specificity in mutually beneficial symbiosis is a significant, yet largely unexplored, prospect of evolutionary biology. These selective constraints can be explored through the study of nucleotide polymorphism at loci controlling specificity. The membrane-anchored receptor *NORK* (nodulation receptor kinase) of the legume *Medicago truncatula* controls early steps of root infection by two symbiotic microorganisms: nitrogen-fixing bacteria (rhizobia) and endomycorrhizal fungi (Glomales). We analyzed the diversity of the gene *NORK* by sequencing 4 kilobases in 28 inbred lines sampled from natural populations. We detected 33 polymorphic sites with only one nonsynonymous change. Analysis based on Tajima's *D* and Fay and Wu's *H* summary statistics revealed no departure from the neutral model. We analyzed divergence using sequences from the closely related species *M. coerulea*. The McDonald-Kreitman test indicated a significant excess of nonsynonymous changes contributing to this divergence. Furthermore, maximum-likelihood analysis of a molecular phylogeny of a few legume species indicated that a number of amino acid sites, likely located in the receptor domain of the protein, evolved under the regime of positive selection. Further research should focus on the rate

and direction of molecular coevolution between microorganisms' signaling molecules and legumes' receptors.

**Key words:** Symbiosis — *Medicago truncatula* — Rhizobium — Neutral theory — Positive selection

### Introduction

Analyzing patterns of sequence polymorphism and between-species divergence of a given gene has become a popular way to elucidate the selective forces acting on its function. Indeed, distinct forms of selection produce specific patterns of sequence diversity (Charlesworth et al. 2001). In the neutral theory of molecular evolution proposed by Kimura (1983), purifying selection was assumed to be ubiquitous. His theory envisioned three types of mutations in DNA sequences: neutral (the function remains unaffected), deleterious (eliminated by selection), and beneficial (expected to be rare enough to be neglected). Therefore, only neutral variation should be observed. However, this model neglects other important forms of selection that occur in genomes, for instance, variable selective pressures over time. Some forms of frequency-dependent selection can lead to balanced polymorphisms that are maintained through selection. Such systems include self-incompatibility loci in plants, incompatibility loci in fungi, and perhaps

Sequence data were deposited in the GenBank database under accession nos. AY676428 to AY676457 and AJ884582.

Correspondence to: Stéphane De Mita, INRA Station de Génétique et Amélioration des Plantes, domaine de Melgueil 34130 Mauguio, France; email: demita@ensam.inra.fr

MHC in vertebrates, as well as resistance systems in plants. They all entail some form of rare allele advantage (Richman 2000; Van der Hoorn et al. 2002). Rare allele advantage prevents the loss of alleles by genetic drift, leading to the maintenance of sometimes very old alleles in populations. This causes patterns such as an excess of polymorphisms segregating at intermediate frequencies and trans-specific polymorphisms (Takahata 1990).

Another important departure from the neutral theory happens when gene products undergo a variable or fast rate of amino acid changes through positive Darwinian selection. At the population level, such directional selection consists in selective sweeps of favorable mutations, inducing hitchhiking effects around the locus under selection (Barton 2000). If enough sweeps occur, positive selection can increase the number of nonsynonymous substitutions between species. An excess of amino acid replacements forms the basis of many methods currently used to document positive selection in genomes (Yang and Bielawski 2000). Positive selection may be more ubiquitous than previously described in the literature. However, its effects are likely to be episodic and restricted to a limited fraction of positions in the protein sequence and, therefore, not easily detectable.

Interspecific interactions are probably a major ecological factor causing nonneutral evolution. Particularly, we expect strong selective pressures on genes involved in recognition mechanisms in host-pathogen relationships (Baum et al. 2002). The evolutionary history of the gene is often shaped by the necessity either to recognize partners (in order to trigger defense or infection) or to escape recognition (to avoid infection or defense mechanisms). This should lead to patterns of fast evolution.

Specific signaling has also been involved in mutually beneficial interactions. Signals may be involved in the localization and/or the recognition of symbiotic partners. It is currently not clear what form of selection is expected in this case. A change in the signaling molecule might be detrimental, so that purifying selection should constrain a gene encoding such a molecule, thereby maintaining a high specificity of recognition. Results compatible with this view have been reported (Jiggins et al. 2002). However, some parasites of the mutualism could be selected to counterfeit compatible symbionts' signals, eliciting an advantage to rare or new specificity signals, leading respectively to balancing and positive selection episodes in both partners. Very few empirical data are available to test these ideas.

Our model interaction is the symbiotic fixation of nitrogen in legumes (Fabaceae), in which the plant hosts bacteria of the genera *Rhizobium*, *Mesorhizobium*, *Bradyrhizobium*, and *Sinorhizobium*

(among others) in specific root nodules. A complex molecular dialogue between plants and bacteria precedes nodule formation, and only compatible bacterial strains can infect roots (van Rhijn and Vanderleyden 1995; Perret et al. 2000). Several bacterial signals have been described, of which the most important are lipo-chito-oligosaccharides (LCOs), also (and hereafter) called Nod factors. It seems likely that plants' specific receptors for Nod factors and other signals allow infection of root tissue by specific compatible strains. Hence, genes encoding these receptors should reflect selective events related to the evolution of the recognition mechanism.

The *Medicago truncatula* gene *NORK* (for *Nodulation Receptor Kinase*) encodes one of the receptors activated early in the infection process (Endre et al. 2002; Stracke et al. 2002). The protein product belongs to the receptor-like kinase class with three leucine-rich repeats (LRRs). Receptor-like kinases are membrane-spanning proteins with an extracellular domain capable of binding a specific ligand and an intracellular kinase domain capable of activating an effector. The LRR motifs are exposed on the extracellular domain and are the active site of this domain (Jones and Jones 1997). Each LRR forms a  $\beta$ -sheet loop structure which allows  $\beta$ -sheets from successive LRRs to be aligned. This structure allows specific binding to a peptide target sequence, but to date the ligand of *NORK* is unknown.

The purpose of this study is to characterize (1) the nucleotide diversity of *NORK* within *Medicago truncatula* and (2) patterns of molecular divergence, using several interspecific comparisons. We sequenced genotypes originating from natural populations of *Medicago truncatula* located throughout the Mediterranean region. In order to study patterns of divergence, we used published expressed sequence tags (ESTs) from orthologous *NORK* genes in several legume species. Our results suggest that the *NORK* gene has been subject to positive selection. We discuss these findings in the light of the biology of legume-microbe symbiotic interactions.

## Materials and Methods

### *Plant Material and Experimental Procedure*

*Medicago truncatula* (the barrel medic) is an annual, diploid plant and is predominantly selfing (Bonnin et al. 1996). We used 28 *Medicago truncatula* ssp. *truncatula* inbred lines available from the collection at INRA Montpellier, which include samples from the entire species natural distribution, except for the Middle East region (J.-M. Prospero, pers. comm.). The accessions were obtained after one or two generations of selfing of plants grown from seeds collected in the wild. The genotypes we sequenced were therefore expected to be highly homozygous. We chose accessions according to their geographical origin, in order to maximize diversity. The

**Table 1.** Accession number and country of origin of the genotypes sequenced in the study

Taxon	Accession	Country of origin	Sequence accession no.
<i>M. truncatula</i>	Jemalong A17	Unkown	AY676428
<i>M. truncatula</i>	L000736	Algeria	AY676435
<i>M. truncatula</i>	L000734	Algeria	AY676434
<i>M. truncatula</i>	L000543	Algeria	AY676446
<i>M. truncatula</i>	L000538	Algeria	AY676432
<i>M. truncatula</i>	L000542	Algeria	AY676433
<i>M. truncatula</i>	L000529	Cyprus*	AY676445
<i>M. truncatula</i>	L000535	Italy*	AY676453
<i>M. truncatula</i>	L000536	Israel*	AY676454
<i>M. truncatula</i>	L000527	Tunisia*	AY676431
<i>M. truncatula</i>	L000534	Jordan*	AY676441
<i>M. truncatula</i>	L000537	Grece (Crete)	AY676444
<i>M. truncatula</i>	L000648	France	AY676442
<i>M. truncatula</i>	L000651	France	AY676443
<i>M. truncatula</i>	L000531	France	AY676430
<i>M. truncatula</i>	L000530	France	AY676429
<i>M. truncatula</i>	L000549	France	AY676450
<i>M. truncatula</i>	L000551	France	AY676452
<i>M. truncatula</i>	L000550	France	AY676455
<i>M. truncatula</i>	L000552	France (Corsica)	AY676438
<i>M. truncatula</i>	L000553	France (Corsica)	AY676437
<i>M. truncatula</i>	L000554	France (Corsica)	AY676451
<i>M. truncatula</i>	L000557	Greece	AY676440
<i>M. truncatula</i>	L000555	Greece	AY676439
<i>M. truncatula</i>	L000547	Spain	AY676448
<i>M. truncatula</i>	L000544	Spain	AY676436
<i>M. truncatula</i>	L000545	Spain	AY676447
<i>M. truncatula</i>	L000548	Spain	AY676449
<i>M. truncatula</i> ssp. <i>tricycla</i>	L000540	Algeria	AY676456
<i>M. littoralis</i>	L000558	France	AY676457
<i>M. coerulea</i>	PI314275	Uzbekistan	AJ884582

*Note.* Lines are available through indicated codes from the collection maintained at UMR “Diversité et génome des plantes cultivées,” Montpellier, France (J.-M. Prosper, pers. comm.). Lines were obtained from natural populations except the lines marked with an asterisk, which were obtained from Australian cultivars that were selected from material collected in the country stated in the table. *M. coerulea* PI314275 consists of nonfixed genetic material available through the National Plant Germplasm System of USDA.

accession codes and geographical origins of each line are given in Table 1. We used three outgroup genotypes, from *M. truncatula* ssp. *tricycla*, *M. littoralis*, and *M. coerulea*. The first two genotypes are very close relatives of *M. truncatula* ssp. *truncatula* and therefore are suitable for orienting polymorphisms (see below), while *M. coerulea*, a more distantly related species, is more useful for interspecific comparisons and the contrast of within-species polymorphism vs. between-species divergence. DNA was extracted from fresh leaves using a protocol adapted from Tai and Tanksley (1990).

We used the published sequence of the *Medicago truncatula* line A17 *NORK* gene, available in DDBJ/EMBL/GenBank (accession number AJ418370), in order to design PCR primers. This sequence is 8569 nucleotides long, with 15 exons. We amplified approximately 4000 nucleotides from the different protein domains: extracellular, LRRs, transmembrane, and kinase, including coding and noncoding sequences (see Fig. 1). These fragments were amplified by PCR, using the following primer pairs at different annealing temperatures: (1) TCGATCGGGGTAACAGAAGT and (2) GATCCAGATGCCTTGACTAA, at 60°C; (3) TGAA-GAGACCAACCAAAAAG and (4) TTCCAACAGCCAAAAGTAATC, at 57°C; and (5) TTTGGACAAGTATTCGTGAT and (6) ATTGGACATGAAAGGATACA, 57°C. See Fig. 1 for primers positions.

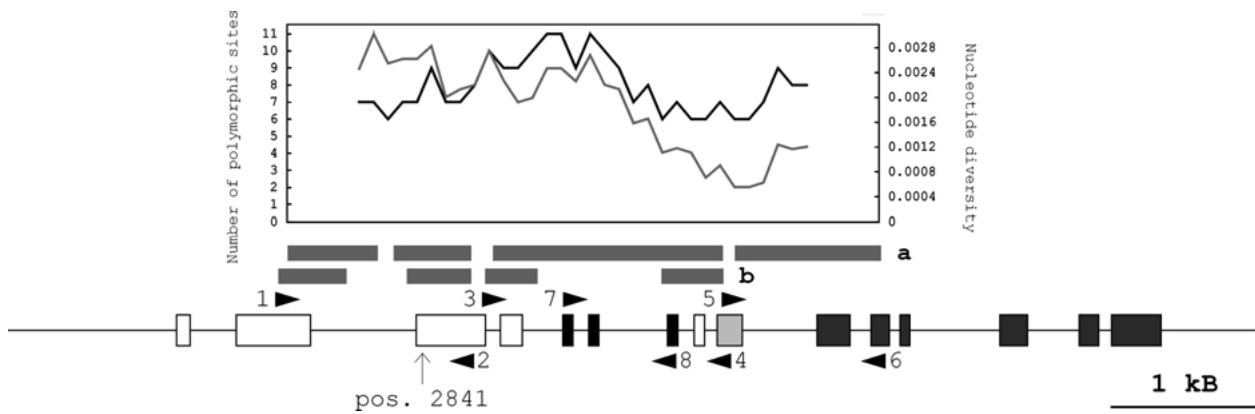
In addition, we amplified and sequenced two intron fragments from other genes of *M. truncatula*. They lie in the *pectate lyase-like* homolog gene (*PLH*; Wu et al. 1996) and the *NADH-dependent*

*glutamate synthase* locus (*GS*; Vance et al. 1995). PCR and sequencing primers were ACCTAATCCTGAAAACATC and GCCATACATCTTGCTCTGC for *PLH* (annealing temperature, 56°C) and AACGGGGAAATCAACACAC and CGAGGACA CCATCAAAAACA for *GS* (annealing temperature, 56°C).

Amplification reactions were performed in a final volume of 25 µl in the presence of 20 ng of template DNA, 10 pmol of each primer, a 0.2 mM concentration of each deoxynucleotide, 1.5 mM MgCl<sub>2</sub>, and 1 unit of Taq polymerase (Sigma). PCR was carried out using a PTC 100 thermocycler (MJ Research). After 5 min at 94°C, 35 cycles were performed of 30 s at 94°C, 30 s at the given annealing temperature depending on the primer pair, and 1 min at 72°C, followed by a final extension step of 5 min at 72°C.

PCR products were purified by gel filtration on Sephadex G50 resin (Amersham Bioscience) and directly sequenced on both strands with the PCR primers using automated sequencing (BigDye Terminator Cycle kit V1.1) on an ABI PRISM 3100. For the LRR fragment, one additional pair of sequencing primers was used ([7] TCTTTCTTCCAATAATCTCA and [8] ATGTGGCAGTGAGA TAATGG).

To estimate sequence divergence, we used EST sequences of *NORK* in *Medicago truncatula*, *Medicago sativa*, *Melilotus alba*, *Vicia hirsuta*, *Pisum sativum*, and *Lotus japonicus* (DDBJ/EMBL/GenBank accession numbers are, respectively, AJ418369, AJ418368, AJ428991, AJ428990, AJ418375, and AF492655). It is very likely that these sequences are orthologous (Endre et al. 2002). We refer to this dataset as “interspecific.”



**Fig. 1.** Structure of the *NORK* locus and sliding window analysis of nucleotide diversity. Boxes indicate translated regions. Colors refer to predicted protein domains: white/black, extracellular; black, LRRs; light gray, transmembrane; dark gray, kinase. Scale: length of 1000 nucleotides (1 kb). Arrowheads above and below exons give the position of primers used in this study. Numbers refer to the primer sequence in Materials and Methods (note that arrowhead tails are at the location where primers start, but arrowheads are longer than the primers). Lines above the gene

indicate sequenced regions. Region a: intraspecific dataset (*M. truncatula* and outgroups *M. truncatula* ssp. *tricycla* and *M. littoralis*). Region b: region sequenced in *M. coerulea*. The top frame presents a sliding window analysis of polymorphism in the intraspecific data (region a). The number of polymorphic sites (black curve) and the nucleotide diversity (gray curve) were computed using a sliding window of length 1000 nucleotides with a step of 100 nucleotides. The vertical arrow indicates the position of the amino acid polymorphism (position 2841).

### Alignment and Detection of Polymorphism

We aligned our *Medicago truncatula* sequences (the intraspecific dataset) using the Genalys software (<http://software.cng.fr/>) and located all polymorphic nucleotide positions in the alignment. Several ambiguities and singleton mutations at polymorphic positions were resolved by resequencing. For each polymorphic nucleotide position, we computed frequencies of each variant nucleotide. Orientation of all sites, which is required for some of the analyses described below, was inferred by parsimony using *M. truncatula* ssp. *tricycla* and *M. littoralis*, which are very close relatives, as outgroups.

We manually built an unrooted genealogical tree connecting all sample haplotype sequences, under the assumption of the infinite-site mutation model and no recombination (Griffiths and Tavaré 1995). In this representation, haplotypes are mapped on the vertices of the tree and all mutations are mapped on the edges (Fig. 2). This method is strongly dependent on assuming no recombination in the history of the sample. To support our assumption, we computed linkage disequilibrium between all pairs of polymorphic sites (measured as  $r^2$ , Pearson's correlation coefficient) using DnaSP software version 3 (Rozas et al. 2003). Under the hypothesis of recombination within a gene sequence, linkage disequilibrium between polymorphic sites should be negatively correlated with physical distance. Physical distance between polymorphic sites was measured using the *M. truncatula* line A17 sequence as a reference. A more rigorous approach is to estimate the scaled recombination rate,  $\rho = 4N_e c$  where  $N_e$  is the effective population size and  $c$  the per nucleotide recombination rate. We estimated  $\rho$  and tested the hypothesis  $\rho = 0$  using the coalescent-based, composite maximum-likelihood estimation procedure of McVean et al. (2002).

### Statistical Tests for Detecting Signatures of Selection

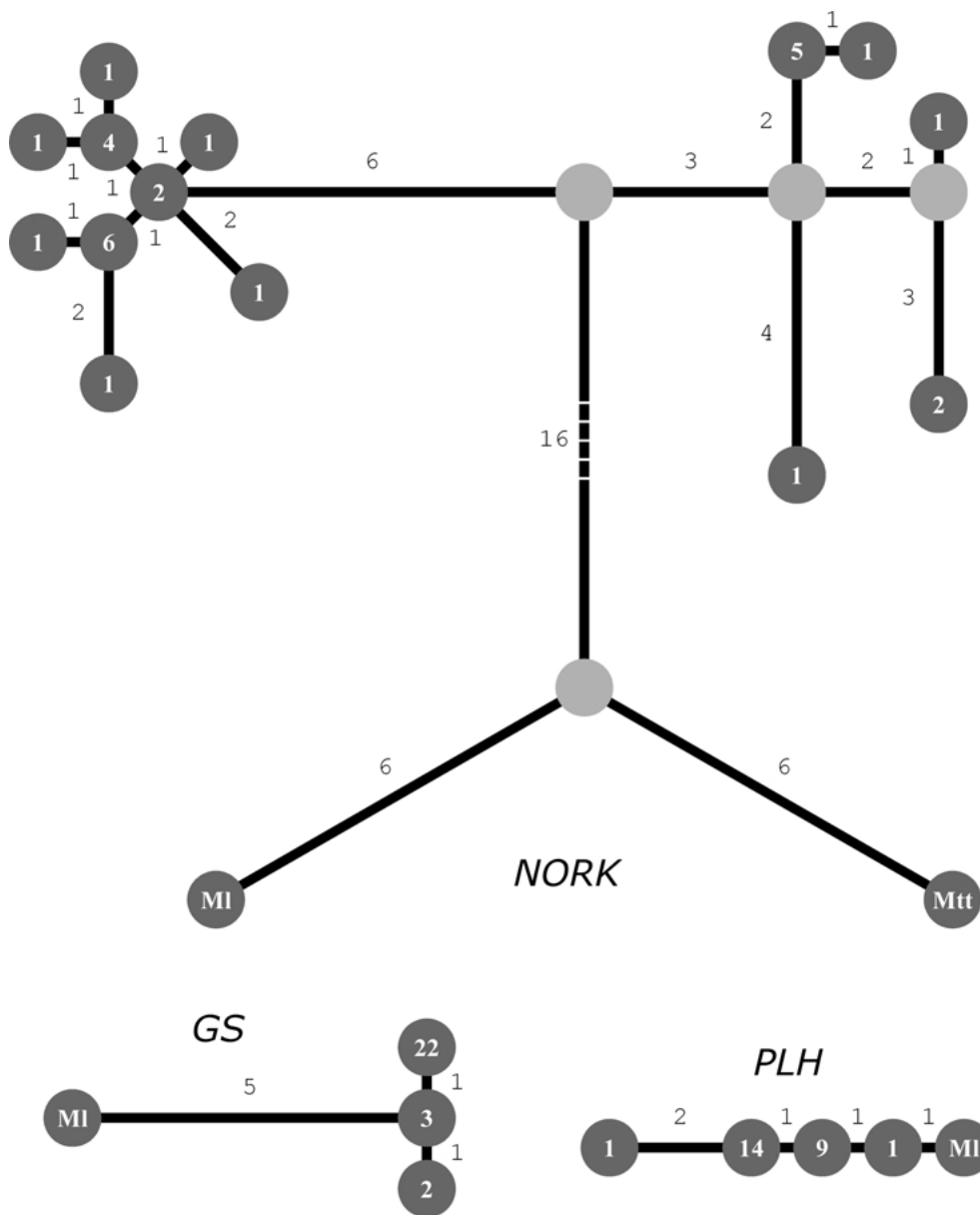
Tajima's (1989)  $D$  and Fay and Wu's (2000)  $H$  are statistics based on the difference between two estimators of the population mutation rate  $\theta = 4N_e \cdot \mu$ , where  $N_e$  is the effective population size and  $\mu$  the mutation rate per nucleotide. Both  $D$  and  $H$  have an expected value close to zero under the hypothesis that the polymorphism is

segregating at equilibrium between neutral mutation and drift.  $D$  was computed using DnaSP and  $H$  was computed manually.

We performed coalescent simulations using the software ms (Hudson 2002) to obtain the distribution of both  $D$  and  $H$  under the neutral model without recombination. Simulations were conditioned on the number of polymorphic sites (see Wall and Hudson 2001). For both tests, the observed value was compared to the simulated distribution under the null hypothesis. We computed the unilateral test probability as the proportion of simulated values more extreme than the observed one. Both the  $D$  and the  $H$  tests were applied to all three loci analyzed (*NORK*, *PLH*, and *GS*).

We applied the McDonald-Kreitman (1991) test to *NORK*, using all polymorphic sites within *M. truncatula*, and substitutions between *M. truncatula* and *M. coerulea*. Due to the amount of polymorphism and divergence available we chose not to perform this test on separate portions of the sequence.

Finally, we analyzed patterns of sequence divergence at *NORK* in the interspecific dataset using codon substitution models (Goldman and Yang 1994). The transition matrix between codons governing the substitution process is controlled by the parameter  $\omega$ , expressed as the ratio between nonsynonymous (dN) and synonymous (dS) substitution rates. We assume that only positive selection can cause a  $dN/dS > 1$  by triggering a great number of adaptive amino acid changes. We used two models of the distribution of  $\omega$  across amino acid sites developed by Yang et al. (2000), which are implemented in the codeml software from the PAML package version 3.14 (Yang 1997). M7 assumes that  $\omega$  is  $\beta$  distributed along the sequence, with a range limited by 0 and 1. Hence M7 does not allow for positive selection while allowing for different intensities of purifying selection. M7 is nested in M8, which assumes a  $\beta$  distribution of  $\omega$  plus a free supplementary category of  $\omega$  value. In the current release of PAML (version 3.14), the supplementary category in M8 is constrained to be  $> 1$ . Therefore, M8 assumes positive selection at some sites in the sequence and the likelihood ratio test (hereafter LRT) of M8 against M7 provides a statistical test for positive selection. We also used an empirical Bayes approach implemented in PAML (Yang et al. 2005) to compute the probability that each site is belonging to each  $\omega$  category of a given model. This probability is a function of the parameters of the model, the frequencies of all categories, and their corresponding  $\omega$  values, which were estimated when fitting the



**Fig. 2.** Haplotype structure of the sample at three loci, represented by unrooted genealogical trees. Haplotypes are represented by dark gray circles (with indication of their frequency in the sample). Haplotypes that are necessary to build the tree, but that were not found in the sample, are shown as open, light gray circles. Outgroup haplotypes are denoted MI (*M. littoralis*) and Mtt (*M.*

*truncatula* ssp. *tricycla*). The numbers on the edges of the tree give the number of mutation steps between haplotypes, excluding three insertion-deletion events and one triallelic site. The asterisk indicates the edge where the only nonsynonymous mutation occurred. The length of each edge is proportional to the number of mutations on the edge.

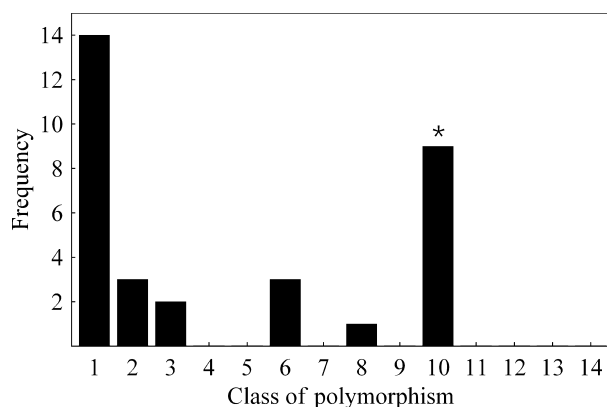
model. This estimation takes into account the uncertainty of maximum-likelihood estimates.

## Results

### *Patterns of Sequence Diversity*

We obtained the sequence of three fragments of *NORK* for 28 inbred lines of *Medicago truncatula*. Overall, we analyzed 3793 sites (Fig. 1). We found three insertion/deletion events, involving, respectively, 1, 2, and 59 nucleotides, and 33 point variants.

All polymorphic sites but one are diallelic in the sample (the last was triallelic). Given the unrooted genealogical tree, all diallelic sites data can be explained by single mutational steps, i.e., there is no homoplasy, whereas three steps are necessary to explain the triallelic site. The distribution of allele frequencies at the 32 diallelic sites is shown in Fig. 3. Only one of these is nonsynonymous. The two states of this amino acid polymorphism are alanine (hydrophobic) and glutamate (acidic). For the *GS* intron fragment, we obtained 27 sequences of 496 nucleotides, containing two diallelic point substitu-



**Fig. 3.** Allele frequency spectrum at locus *NORK*. The plot represents allele frequencies of 32 diallelic polymorphisms (excluding three insertion-deletion events and one triallelic site). Classes of polymorphism (singleton, doubleton, etc.) are represented on the *X* axis. The frequency of each class is represented on the *Y* axis. Note that only the frequency of the rarer allele for each site is represented. The only nonsynonymous mutation found in our sample is indicated here by an asterisk.

tions. For the *PLH* intron fragment, we obtained 25 sequences of 277 nucleotides, containing four diallelic point substitutions, and an 8-nucleotide-long region showing two insertion/deletion events plus a diallelic point substitution.  $r^2$  does not decrease with physical distance, suggesting an absence of recombination in the history of the sample at all three loci. Accordingly, the estimated population recombination rate  $\rho$  is zero for each locus.

The patterns of sequence diversity we observed in *M. truncatula* at the three loci are summarized in Table 2. Tajima's  $D$  is positive and Fay and Wu's  $H$  is positive at *NORK*. Both  $D$  and  $H$  are negative at *PLH*. *GS* displays a negative  $D$  and a positive  $H$ . None of these tests is significant at the 5% level. Note that the very low polymorphism at *GS* and *PLH* severely limits the power of both tests to detect departure from neutrality.

The pattern of diversity along *NORK* was examined through a sliding window analysis displayed in Fig. 1. Note that there is probably too few polymorphic sites to reliably estimate local variation. However, the nucleotide diversity is clearly higher in the 5' region than in the 3' region, whereas the number of polymorphic sites is roughly constant across regions. This trend suggests that the allelic frequencies are more balanced (and thus Tajima's  $D$  is more positive) in the 5' region than in the 3' region.

#### McDonald-Kreitman Test

We estimated fixed sites divergence between *M. truncatula* and *M. coerulea*. There are 34 such sites, among the 1636 sites we examined (Fig 1), and 21 of them are synonymous changes. The McDonald-

Kreitman test is highly significant ( $p < 0.001$ ; Table 3) and reflects either an excess of nonsynonymous interspecific substitutions (through positive selection) or a deficit of intraspecific nonsynonymous mutations (caused by intense purifying selection in *M. truncatula* clade since its divergence from *M. coerulea*).

#### Codon Substitution Models

To further investigate the possibility of positive selection, we performed a maximum-likelihood-based analysis of codon substitution models. This analysis was performed on a sequence alignment spanning 926 codons in six different legume species. Taking only point mutations into account, 468 codons are variable. The underlying species tree is shown in Fig. 4. The parameters of two models of coding sequence evolution, taking into account the variation of  $\omega$ , were estimated. We tested the hypothesis of positive selection against purifying selection with a LRT (M8 versus M7 model). The M8 model, which imposes positive selection, fits the data much better than the M7 model (Table 4). M7's  $\omega$  categories are distributed between 0 and 0.98, presumably to fit the occurrence of some neutrally evolving or positively selected sites in the data. Estimation of parameters in model M8 suggests that 0.15 of codons (139) fall in a category with estimated  $\omega = 1.31$ . The other codons fall into  $\omega$  categories with values between 0 and 0.20, i.e., a more restricted parameter space than the M7 estimates. To address the problem of spurious detection of positive selection, we performed the test of model M8 against a version of M8 wherein the supplementary category  $\omega$  is fixed at 1 (model M8A; Swanson et al. 2003). This test is nonsignificant, suggesting that the signature of positive selection is not strong enough to be detected by this test. We argue about this point in the Discussion. The empirical Bayes procedure allowed us to estimate, for each site, the probability of belonging to each  $\omega$  category as a function of the maximum-likelihood estimates of the parameters obtained under model M8. We could then compute a "predicted  $\omega$ " for each site, as the weighted mean of  $\omega$  categories. An estimate of the sampling variance around the predicted  $\omega$  is also provided (Fig. 5). Table 5 gives the number of sites for which the estimated  $\omega$  was  $> 1$  (hereafter "candidate" sites). There is a highly nonrandom distribution of the proportion of candidate sites between protein domains (Table 5), with significantly more candidate sites in the solvent-exposed regions of *NORK* (extracellular domain) than in the other regions (signal peptide, transmembrane and kinase domains; exact test,  $p < 0.0001$ ). The excess of candidate sites is striking for the LRR region, suggesting

**Table 2.** Summary of nucleotide diversity

Locus	<i>NORK</i>	<i>PLH</i>	<i>GS</i>
Number of sequences used ( <i>n</i> )	28	25	27
Length of the sequenced region (nucleotides)	3793	277	496
Number of polymorphic (segregating) sites	33	4	2
Nucleotide diversity ( $\pi$ ), per site	0.00220	0.00312	0.00092
Watterson's estimator of $\theta$ , per site	0.00227	0.00447	0.00105
Tajima's <i>D</i>	-0.11	-0.81	-0.26
<i>p</i> value for Tajima's <i>D</i>	0.51	0.26	0.42
Fay and Wu's <i>H</i>	1.06	-0.57	0.60
<i>p</i> value for Fay and Wu's <i>H</i>	0.15	0.33	0.24

**Table 3.** McDonald-Kreitman test at the *NORK* locus

	Polymorphism	Divergence
Synonymous mutations	31	21
Nonsynonymous mutations	1	13

*Note.* We took into account only diallelic point substitutions.

that positive selection affected this region. There are also comparatively too few candidate sites in the transmembrane and kinase domains, although the test is not significant for the transmembrane domain, most probably because of a lack of power. Additionally, candidate sites within the LRR motifs seem to be localized downstream of each repeat, that is, after the predicted  $\beta$ -sheet structural domain (Fig. 5).

## Discussion

### *Molecular Signatures of Natural Selection*

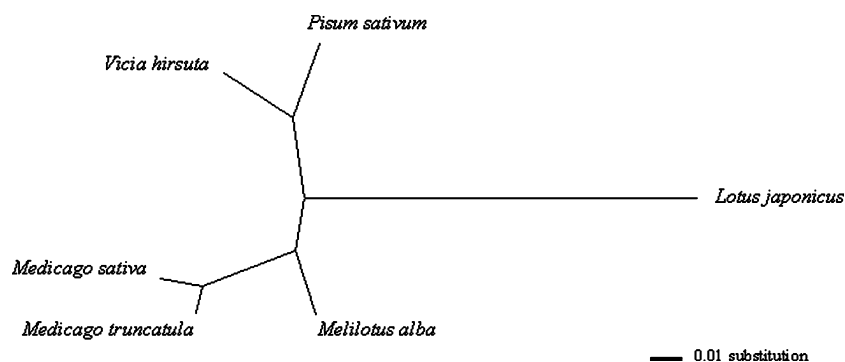
The neutral, constant-sized model can not be rejected for the *NORK* locus, or for the two control loci. Note that tests at control loci are based on low amounts of polymorphism and thus had very little power to reject the null model. Within-sample subdivision and the self-fertilizing nature of *M. truncatula* complicate tests of neutrality, because ideally a neutral model incorporating both features should be used.

We found only one nonsynonymous mutation in the sample, indicating that fairly strong purifying selection is operating on the *NORK* protein. Interestingly, this site segregates at an intermediate frequency (Fig. 3) and codes for amino acids from distinct classes: the ancestral one is the hydrophobic alanine, and the derived one is the acidic glutamate. At this position, an alanine is found in all other legume sequences used for the interspecific analysis. The codon-based analysis classifies this site as under purifying selection (estimated  $\omega = 0.4 \pm 0.4$ ), suggesting that it is functionally important. These findings are consistent with a putative adaptive role of the mutation found in *M. truncatula*, affecting, for instance, the structural conformation of the protein

or its affinity to a ligand. But more evidence is needed before concluding the status of that mutation.

By contrast, the interspecific data provide clear evidence for positive selection. The McDonald-Kreitman test (Table 3) suggests an excess of nonsynonymous substitutions during the divergence between *M. truncatula* and *M. coerulea* (a diploid species which is closely related to the tetraploid alfalfa *M. sativa*). The observed number of amino acid changes is highly unlikely under neutrality, but several factors not necessarily involving positive selection might result in a relative excess of nonsynonymous changes. A drastic reduction in effective population size can modify the intensity of genetic drift and thus the efficiency of selection against deleterious mutations (Eyre-Walker 2002). In addition, the mating system has changed several times between selfing and outcrossing in the history of the genus *Medicago* (Béna et al. 1998), particularly between *M. truncatula* (selfer) and *M. coerulea* (outcrosser), and high levels of selfing can halve the effective population size relative to an outcrosser (Nordborg 2001). However, these factors alone are not sufficient to explain the observed pattern. First, although a reduction of population size can increase the rate of nonsynonymous substitutions, it is not likely to exceed the synonymous rate. Yet the codon substitution model suggests that several sites have predicted  $\omega$  that are significantly bigger than 1. Second, we detected significant variation in evolutionary constraints between protein domains (Table 5), which cannot be easily explained by nonselective hypotheses alone.

Finally, the use of EST data requires caution, as methods used in EST sequencing may not be as accurate as the sequencing protocol we used for surveying sequence polymorphism. However, these errors will involve equally often nonsynonymous and synonymous changes, biasing  $dN/dS$  toward 1. Some sites may thus be detected as false positives even in absence of positive selection. But such sites should occur randomly with respect to their position on the gene sequence. In our case the significant variation of  $\omega$  estimates along *NORK* supports the hypothesis



**Fig. 4.** Neighbor-joining tree of *Medicago truncatula*, *Medicago sativa*, *Melilotus alba*, *Vicia hirsuta*, *Pisum sativum*, and *Lotus japonicus* of *NORK* EST sequences.

**Table 4.** Comparison of the log-likelihood of codon substitution models

Model	$p$	$\beta$ parameters	$\omega_s$	$p_s$	$\log L$	LRT
M7 ( $\beta$ )	2	0.18, 0.48	NE	NE	-7228.84	5.36* (vs. M7)
M8A ( $\beta + \omega = 1$ )	3	9.64, 99	1	0.21	-7226.16	1.85 (ns) (vs. M8A)
M8 ( $\beta + \omega > 1$ )	4	12.58, 99	1.31	0.15	-7225.23	7.22* (vs. M7)

*Note.*  $p$ : number of parameters fitted in the model. NE: not estimated in the model.  $\omega_s$ : maximum likelihood estimate of the value of  $\omega$  for the extra category of sites.  $p_s$ : maximum likelihood estimate proportion of codon evolving at the rate  $\omega_s$ . LRT: likelihood ratio test statistic. The asterisk indicates that the probability of observing such an LRT or higher under the M7 or M8A model is  $< 0.05$ , assuming that the LRT follows a  $\chi^2$  distribution with 2 degrees of freedom (for the test of M8 vs. M7) and 1 degree of freedom (for the test of M8 vs. M8A). ns: nonsignificant.

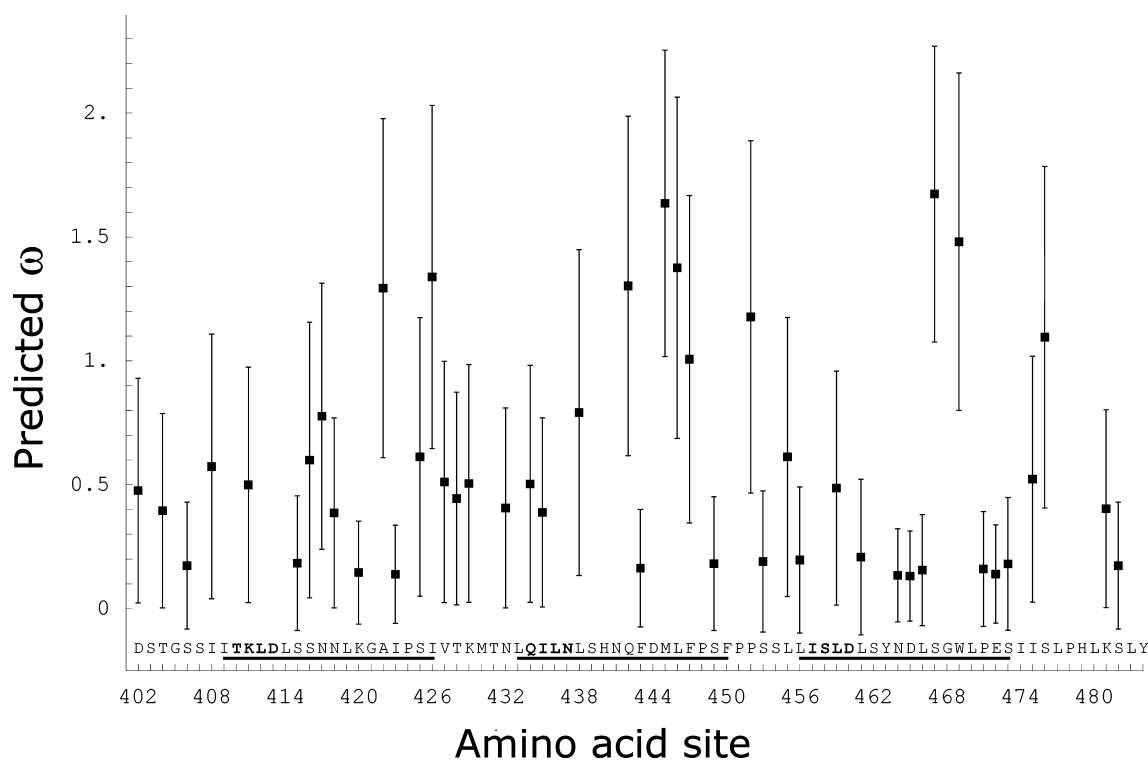
of positive selection (Table 5), and we conclude that positive selection must shape some of the patterns of molecular divergence among the legume species studied here.

Recently, a few studies have reported the possibility of spurious detection of positive selection using the framework we used (Bierne and Eyre-Walker 2003; Swanson et al. 2003; Zhang 2004). Zhang (2004) reported frequent false detection of positive selection but he focused on models allowing for branch-specific heterogeneity of selective forces (these were not used here). Simulations results of Anisimova et al. (2001) suggest fair rates of false positive (5%) and a fairly good power (77%) when considering simulation parameters fitting our own data. However, simulations in their power analysis assumed a rather strong positive selection ( $\omega > 2$ ). Wong et al. (2004) studied several other cases. Considering the simulation schemes best fitting our data (schemes 2b and 5 with 5 taxa), false positives did not exceed 5% and the rate of true positives (power) was quite low. More precisely, the power of the M7-vs.-M8 comparison was fairly high (52%) but the M8A-vs.-M8 test's power was strongly reduced (20–28%). Therefore, simulation studies confirm that the type I error rate for our M7-vs.-M8 test is not likely to be inflated (at the 5% level) but suggest a rather poor power, especially for the more stringent M8A-vs.-M8 test. Given that the potential signal of positive selection in our data is probably rather weak (the supplementary  $\omega$  value estimated in model M8 is only 1.31), we argue

that in our case the test of M8A vs. M8 probably has a very low power for testing explicitly the hypothesis  $\omega > 1$ . We ran additional analysis under the former release of PAML (3.14 beta 3), where  $\omega$  in M8 is not constrained to be  $> 1$  (data not shown). We chose a low starting  $\omega$  value (0.01), in order to check that the occurrence of sites with  $\omega > 1$  was not an artifact. Results were identical than those presented, indicating that the positive selection signal was not induced by the  $\omega > 1$  assumption. By contrast, when analyzing only the transmembrane-kinase portion of coding sequences, the supplementary  $\omega$  category was either null (frequency 0) or redundant with a category generated by the  $\beta$  distribution, depending on the starting value. In both cases, the M7-vs.-M8 likelihood ratio test was not significant. Eventually, only an analysis comprising many more taxa (that would decrease type I error and increase the statistical power of tests) may definitively resolve this issue.

#### Localization of Selection Footprints

An estimate of per-site  $\omega$  ratios in the LRR region is shown in Fig. 5. This estimation was based on individual site data and the sampling variance of the  $\omega$  individual estimates is rather large, as shown by the standard errors. Thus, this method is more relevant for a general description than for a detailed, site-specific analysis, because the rates of false positives and false negatives are not estimated.



**Fig. 5.** Predicted  $\omega$  at individual amino acid sites of the LRR region of the *NORK* gene with standard error estimates. Only data from variable sites between site 402 and site 484 are represented. *Medicago truncatula* amino acid residues are indicated along the *X* axis. Amino acid positions of LRRs are underlined, and residues in boldface indicate the positions of  $\beta$ -sheets.

**Table 5.** Distribution of sites with different predicted  $\omega$  among *NORK* domains

Protein domain	Boundaries	Number of sites	Sites with $\omega < 1$	Sites with $\omega > 1$	Expected number of sites with $\omega > 1$	<i>p</i> value
Signal peptide	1–30	30	26	4	2	0.08
Extracellular	31–407	377	352	25	21	0.21
LRRs	408–478	71	61	10	4	< 0.01
Degenerated LRRs	479–520	42	37	5	2	0.08
Transmembrane	521–543	23	23	0	1	0.64
Protein kinase	544–926	383	376	7	21	< 0.001
Total		926	875	51	51	

*Note.* Boundaries (in nucleotide positions) are based on Stracke et al. (2002). The expectation of the number of sites with  $\omega > 1$  is computed as the product of marginal values, rounded off to the nearest integer. An overall test of independency of the table is significant (Fisher's exact test,  $p < 0.0001$ ). Domain *p* values are obtained from binomial exact tests (LRRs and kinase domains remain significant at the 5% level even when applying a Bonferroni correction for multiple testing).

Our analysis suggests that the sites under positive selection are found mainly in the extracellular domain, including the LRR region (Table 5). The ligand binding specificity of LRR receptor-like kinases depends on the amino acid sequence of the LRR region (Jones and Jones 1997) but is also probably influenced by other positions of the extracellular domain (Luck et al. 2000). Bergelson et al. (2001) reviewed positive selection in proteins belonging to the nucleotide-binding-site-LRR family involved in pathogen resistance. The strongest footprints of positive selection were often concentrated in LRR regions, as in

the case of *NORK*. However, half of the 51 candidate sites of our study were found in the 5' end of the extracellular domain (Table 5), indicating that this region may also play an active role in ligand recognition. We found, in *NORK*'s LRR region, 10 sites that may be under positive selection (Fig. 5). Those sites are at positions 422(C), 426(R), 442(A), 445(N), 446(Y), 447(T), 452(T), 467(S), 469(R), and 476(F), the letter indicating the amino acid found in the *M. truncatula* sequence. Among them, sites 445 and 467 are the best candidates. These sites are located in the second half of each LRR motif, indicating that

positive selection occurred in regions immediately downstream of the  $\beta$ -sheet structures. These sites should be viewed as suggestive instances of positive selection but whether individual sites are false positives or have experienced genuine adaptive evolution cannot be explicitly addressed here.

#### *Implications for the Evolution of Legume-Microbe Interactions*

Our results suggest an episode of adaptive evolution of NORK in the history of the legume family. Its key role in rhizobial and mycorrhizal interactions suggests that symbiotic selective forces may have triggered rapid adaptive changes in this lineage. Scenarios involving change in specificity, such as rapid coevolution or symbiont shifting, could be invoked. However, our data provide no way to distinguish this proposal against other selection pressures. For instance, symbiotic receptors can be used by parasites as an entry route, provided that they exhibit molecules analogous to symbiotic signals. Then, strong selective pressure for parasite resistance may induce patterns of positive selection at genes involved, without a direct connection to the evolution of mutualism. Addressing this point requires other kinds of analysis than those presented here.

Previous work has uncovered several examples of either positive or balancing selection on genes involved in host-pathogen relationships. This is straightforwardly explained by strong selective pressures imposed by each partner. Our results show that rapid adaptive amino acid changes might also be triggered in mutualistic interactions. One may now ask whether coevolution (that is, reciprocal adaptation) occurs between plants and symbiotic microbes. To address this question we need to know by which mechanism the NORK receptor is activated and, thus, which symbiotic signal molecules are involved. In that respect, recent research has revealed several other receptor-like kinase-encoding genes required for Nod factor perception (Limpens et al. 2003; Radutoiu et al. 2003). These genes encode a Nod factor receptor and are involved only in the nodulation process. Therefore they are probably good candidates to seek more specifically footprints of selection due to coevolution in legume-rhizobium recognition systems.

*Acknowledgments.* We thank Mikkel Schierup, Leif Schauer, Nicolas Galtier, Deborah Charlesworth, and two anonymous reviewers for useful comments on early versions of the manuscript, Nicolas Galtier for helpful discussion, and Ziheng Yang, Wendy Wong, and Rasmus Nielsen for communicating the manuscript before publication.

#### References

- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18:1585–1592
- Barton NH (2000) Genetic hitchhiking. *Phil Trans R Soc Lond B* 355:1553–1562
- Baum J, Ward RH, Conway DJ (2002) Natural selection on the erythrocyte surface. *Mol Biol Evol* 19:223–229
- Béna G, Lejeune B, Prosperi J-M, Olivieri I (1998) Molecular phylogenetic approach for studying life-history evolution: the ambiguous example of the genus *Medicago* L. *Proc R Soc Lond B* 265:1141–1151
- Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant R-genes. *Science* 292:2281–2285
- Bierne N, Eyre-Walker A (2003) The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* 165:1587–1597
- Bonnin I, Huguët T, Gherardi M, Prosperi J-M, Olivieri I (1996) High level of polymorphism and spatial structure in a selfing plant species, *Medicago truncatula* (Leguminosae), shown using RAPD markers. *Am J Bot* 83:843–855
- Charlesworth D, Charlesworth B, McVean GAT (2001) Genome sequences and evolutionary biology, a two-way interaction. *Trends Ecol Evol* 16:235–242
- Endre G, Kereszt A, Kevei Z, Mihacea S, Kalò P, Kiss GB (2002) A receptor kinase gene regulating symbiotic nodule development. *Nature* 417:962–966
- Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics* 162:2017–2024
- Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413
- Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736
- Griffiths RC, Tavaré S (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math Biosci* 127:77–98
- Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338
- Jiggins FM, Hurst GDD, Yang Z (2002) Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae. *Mol Biol Evol* 19:1341–1349
- Jones DA, Jones JDG (1997) The role of leucine-rich repeat proteins in plant defences. *Adv Bot Res Adv Plant Pathol* 24:89–167
- Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge
- Limpens E, Franken C, Smit P, Willemsse J, Bisseling T, Geurts R (2003) LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. *Science* 302:630–633
- Luck JE, Lawrence GJ, Dodds PN, Shepherd KW, Ellis JG (2000) Regions outside of the leucine-rich repeats of flax rust resistance proteins play a role in specificity determination. *Plant Cell* 12:1367–1377
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654
- McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241
- Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds) *Handbook of statistical genetics*. John Wiley & Sons, Chichester, UK
- Perret X, Staehelin C, Broughton WJ (2000) Molecular basis of symbiotic promiscuity. *Microbiol Mol Biol Rev* 64:180–201

- Radutoiu S, Madsen LH, Madsen EB, Felle HH, Umehara Y, Gronlund M, Sato S, Nakamura Y, Tabata S, Sandal N, Stougaard J (2003) Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* 425:585–592
- Richman AD (2000) Evolution of balanced genetic polymorphism. *Mol Ecol* 9:1953–1963
- Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497
- Stracke S, Kistner C, Yoshida S, Mulder L, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J, Szczyglowski K, Parniske M (2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* 417:959–962
- Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18–20
- Tai TH, Tanksley SD (1990) A rapid and inexpensive method for isolation of total DNA from dehydrated plant tissue. *Plant Mol Biol Rep* 8:297–303
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Takahata N (1990) A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc Natl Acad Sci USA* 87:2419–2423
- Van der Hoorn RAL, De Wit PJGM, Joosten MHAJ (2002) Balancing selection favors guarding resistance proteins. *Trends Plant Sci* 7:67–71
- van Rhijn P, Vanderleyden J (1995) The *Rhizobium*-plant symbiosis. *Microbiol Rev* 59:124–142
- Vance CP, Miller SS, Gregerson RG, Samac DA, Robinson DL, Gantt JS (1995) Alfalfa *NADH-dependent glutamate synthase*: structure of the gene and importance in symbiotic N<sub>2</sub> fixation. *Plant J* 8:345–358
- Wall JD, Hudson RR (2001) Coalescent simulations and statistical tests of neutrality. *Mol Biol Evol* 18:1134–1135
- Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051
- Wu Y, Qiu X, Du S, Erickson L (1996) *PO149*, a new member of pollen pectate lyase-like gene family from alfalfa. *Plant Mol Biol* 32:1037–1042
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496–503
- Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Zhang J (2004) Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol* 21:1332–1339