the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2010) **63**, 623–635 doi: 10.1111/j.1365-313X.2010.04267.x

Population genomic analysis of Tunisian *Medicago truncatula* reveals candidates for local adaptation

Maren L. Friesen[1,*], Matilde A. Cordeiro[2,3], R. Varma Penmetsa[2], Mounawer Badri[4], Thierry Huguet[5], Mohamed E. Aouani[4,6], Douglas R. Cook[2] and Sergey V. Nuzhdin[1]

[1]*Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA,*
[2]*Department of Plant Pathology, University of California–Davis, Davis, CA, USA,*
[3]*Instituto de Tecnologia Química e Biológica, Portugal,*
[4]*Center of Biotechnology of Borj Cedria, Tunisia,*
[5]*Ecole Nationale Supérieure Agronomique de Toulouse, France, and*
[6]*NEPAD/North Africa Biosciences Network, Cairo, Egypt*

Received 4 March 2010; revised 19 May 2010; accepted 25 May 2010; published online 22 June 2010.
*For correspondence (fax +213 740 8631; e-mail friesen@usc.edu).

SUMMARY

Genome-wide association studies rely upon segregating natural genetic variation, particularly the patterns of polymorphism and correlation between adjacent markers. To facilitate association studies in the model legume *Medicago truncatula*, we present a genome-scale polymorphism scan using existing Affymetrix microarrays. We develop and validate a method that uses a simple information-criteria algorithm to call polymorphism from microarray data without reliance on a reference genotype. We genotype 12 inbred *M. truncatula* lines sampled from four wild Tunisian populations and find polymorphisms at approximately 7% of features, comprising 31 419 probes. Only approximately 3% of these markers assort by population, and of these only 10% differentiate between populations from saline and non-saline sites. Fifty-two differentiated probes with unique genome locations correspond to 18 distinct genome regions. Sanger resequencing was used to characterize a subset of maker loci and develop a single nucleotide polymorphism (SNP)-typing assay that confirmed marker assortment by habitat in an independent sample of 33 individuals from the four populations. Genome-wide linkage disequilibrium (LD) extends on average for approximately 10 kb, falling to background levels by approximately 500 kb. A similar range of LD decay was observed in the 18 genome regions that assort by habitat; these LD blocks delimit candidate genes for local adaptation, many of which encode proteins with predicted functions in abiotic stress tolerance and are targets for functional genomic studies. Tunisian *M. truncatula* populations contain substantial amounts of genetic variation that is structured in relatively small LD blocks, suggesting a history of migration and recombination. These populations provide a strong resource for genome-wide association studies.

Keywords: linkage disequilibrium, *Medicago truncatula*, local adaptation, genomics, salinity, polymorphism.

INTRODUCTION

The genus *Medicago* contains 83 species, including alfalfa (*Medicago sativa*), that are typically either tetraploid perennial or diploid annual species (Lesins and Lesins, 1979; Small and Jomphe, 1989). *Medicago truncatula* is an exemplar of the annual diploid 'Medics' that occur spontaneously throughout the Mediterranean basin across a wide range of habitats. Because *M. truncatula* has been domesticated in Western Australia for use as forage in dry land agriculture, it is able to serve as a reference species for crop legumes, as well as a model species to understand the molecular genetic basis of legume processes. Of particular interest are the mutualistic interactions with nitrogen-fixing rhizobia and symbiotic mycorrhizal fungi (Heath and Tiffin, 2007, 2009; Rangin *et al.*, 2008; Chen *et al.*, 2009; Gomez *et al.*, 2009), properties shared by the majority of legume species. *Medicago* populations occur naturally across a broad range of stressful habitats, including serpentine soils in California, soils contaminated with heavy metals, drought-impacted regions of Mediterranean countries, and naturally occurring saline soils in North Africa and Western Europe. Salinity

is a major factor affecting agricultural production worldwide, with one-fifth to one-third of irrigated agricultural land at risk (Tester and Davenport, 2003). Thus, identifying genes involved in salinity adaptation in the model legume *M. truncatula* holds promise for enabling the improvement of economically important legume crops through translational genomics (Young and Udvardi, 2009).

Phenotypic differences between populations can be due to drift or selection. In wild populations of *Arabidopsis thaliana*, flowering time and major genes known to underlie it vary with latitude in a manner consistent with evolutionary-ecological theory (Aranzana *et al.*, 2005; Zhao *et al.*, 2007; Wilczek *et al.*, 2009). Spatial differences in selection can occur in response to a range of abiotic and biotic factors to produce locally adapted genotypes. For example, wild populations of the native North American grass *Andropogon gerardii* and their arbuscular mycorrhizal fungi are both locally adapted to the physico-chemical properties of their home soils as well as to one another (Johnson *et al.*, 2010). Whole-genome scans can identify candidate genes for local adaptation, as in the case of *Arabidopsis lyrata* populations growing on serpentine and nearby non-serpentine soils (Turner *et al.*, 2010).

The advent of reference genome sequence enables genome-wide scans of polymorphism in a wide range of organisms. Knowledge of genome-wide polymorphism has great utility for association mapping of ecological or agronomic traits of interest. The goal of association mapping is to identify genetic markers and candidate genes that are correlated with particular phenotypes; this can only succeed if enough markers are sampled to capture each linkage disequilibrium (LD) block within the population. Linkage disequilibrium in a given population depends on mating system, mutation, recombination, and migration rates, as well as patterns of selection (Slatkin, 2008). Since individuals mate within their local population, isolated populations can diverge genetically from one another via genetic drift; within-population LD can be quite low but differences in allele frequencies between populations will lead to high global LD in the species via the Wahlund effect (Slatkin, 2008). Population structure can also induce a large number of false positives in genome-wide association studies, since alleles that differ due to drift can become spuriously correlated with phenotypes that differ among populations (Aranzana *et al.*, 2005; Rosenberg and Nordborg, 2006).

Currently, there is little genome-wide information available about population differentiation and LD in *M. truncatula*. Since *M. truncatula* is highly selfing, with estimates from 95 to 99% (Chaulet and Prosperi, 1994; Bonnin *et al.*, 2001; Siol *et al.*, 2008), LD is expected to be high. However, existing studies show that natural populations vary greatly in their genetic diversity, fine-scale spatial structure, and the number of loci that are in LD. Within French populations, the frequency of loci that exhibit significant LD ranges from 6% of 22 random amplified of polymorphic DNA (RAPD) loci (Aude) to 78% of 13 simple sequence repeat (SSR) loci (Salses in 1999); within subpopulations of Aude, only 0–2% of loci show significant LD (Bonnin *et al.*, 1996; Siol *et al.*, 2007). Analysis of 10 Tunisian populations found on average only 6.3% of 18 loci to be in significant LD, with 30.5% in one population but 0% in others (Lazrek *et al.*, 2009), while another study focused on four populations found on average 20.5% of 20 loci to have significant LD (Badri *et al.*, 2007). A range-wide survey of 13 SSRs in 384 *M. truncatula* lines found 37.2% to have significant LD, which could be due to population structure (Ronfort *et al.*, 2006). Finally, sequencing of three regions spanning a symbiosis gene in 28 lines distributed around the Mediterranean showed that LD did not decay over 50 kb (De Mita *et al.*, 2006). However, this study inferred that positive selection may have acted at this locus; if recent, this would cause extended LD due to a selective sweep.

To enable genome-wide association studies in the model legume *M. truncatula*, we require a genome level picture of polymorphism and LD. Here, we use existing Affymetrix microarrays to perform a genome-wide study of natural variation in *M. truncatula*. We obtain data from 12 inbred genotypes sampled from four Tunisian populations, with two of the populations occurring on high-salinity soils. We develop a new algorithm to call single feature polymorphisms (SFPs) based on information criteria and validate our experiment and algorithm using traditional Sanger sequencing. We describe the patterns of: (i) polymorphism, (ii) population differentiation, and (iii) LD. Finally, we identify genetic regions that are consistently differentiated between saline-source and non-saline-source genotypes. Genotyping an additional 33 individuals for six test loci confirms that these loci are significantly differentiated across habitats. These genome regions contain candidate genes for local adaptation to high-salinity habitats, including several genes with putative roles in abiotic stress responses.

RESULTS AND DISCUSSION

Microarray genotyping

We used inbred lines of *M. truncatula* sampled from four wild populations in Tunisia, shown in Figure 1. We hybridized genomic DNA from 12 individual genotypes (three per population) to existing Affymetrix microarrays to perform a genome-wide study of natural variation in this organism. DNA from a single inbred genotype was hybridized to each array so that haplotypes could be inferred. Oligonucleotide microarrays are being used as a cost-effective technique for simultaneously interrogating hundreds of thousands to millions of genome positions simultaneously. Sequence differences lead to altered hybridization intensity of particular probes, termed SFPs. The first of these studies was performed in yeast and

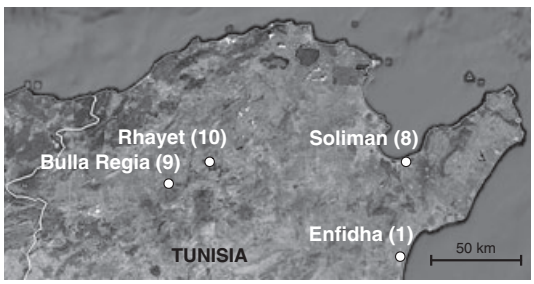

Figure 1. Map showing sampling locations for *Medicago truncatula* germplasm.
Sites 1 (Enfidha) and 8 (Soliman) are highly saline, while sites 9 (Bulla Regia) and 10 (Rhayet) have very low levels of soil salinity.

compared recombinants to their two parental genotypes at the observed 3714 SFPs (Winzeler, 1998). Similarly, in *A. thaliana* two reference lines were hybridized to replicate microarrays to identify 3806 SFPs (Borevitz *et al.*, 2003). A larger sample of 23 accessions uncovered 77 420 SFPs relative to the reference genotype Columbia (Borevitz *et al.*, 2007). However, in natural populations such as those that we are investigating there is no 'reference' hybridization with which to compare hybridization intensity. Studies thus far typically compare between subpopulations or other predefined groups to discover segregating SFPs (Turner *et al.*, 2005, 2008a,b). We extend this approach by developing a new algorithm that uses information criteria to identify probes in our sample that give polymorphic hybridization signal intensities and thus are likely to contain segregating sequence polymorphisms.

Statistical identification of marker probes

As our experiment uses unknown genotypes from natural populations, we do not know a priori the loci where individuals differ from one another. Indeed, since SSR analysis suggests that gene flow is common between populations (Lazrek *et al.*, 2009), identifying loci by their population-level divergence is expected to drastically misestimate the pattern of polymorphism. We need to first determine whether the hybridization intensities observed for each probe reflect one allele or two, then we can calculate LD and determine population-level patterns after assigning individuals' genotypes at loci where we believe two alleles are present. While sophisticated clustering and partitioning algorithms exist, these tend to be computationally expensive, particularly when hundreds of thousands of probes need to be considered. Thus, we develop a simple algorithm based on information criteria (Akaike's an information criterion, AICc; Burnham and Anderson, 2002) to decide whether a two-group model, where the values of each locus are drawn from two distinct distributions, fits the data substantially better than a model where each intensity at a locus is drawn from the same underlying distribution. All computations were performed using R (R Team 2009); code is available upon request.

Polymorphism rate, LD, population structure

At 5% FDR the slide-mean normalized raw data contains 31,419 probes called as markers by our algorithm; these are listed in Table S1 in Supporting Information. Clustering the 12 haplotypes with the unweighted pair group method with arithmetic mean (UPGMA) shows evidence of population structure, since individuals from the same population tend to be more similar to one another, as seen in Figure 2(a). Running 'Structure' (Pritchard *et al.*, 2000) on a subset of the data suggests that five groups best fit the data, but with only 12 individuals we regard this clustering method as highly preliminary since groups contain only two or three individuals (Figure 2b). Nonetheless, 'Structure' provides evidence against the one- and two-population models, which suggests that there is complex population structure in these data. However, despite these indications of population structure it is important to note that the vast majority of SFPs do not assort along population subdivisions: only 938 (3% of probes) are structured by the four populations, and just 90 probes assort with saline habitat. For comparison, 30 probes discriminate populations 1 and 9 from 8 and 10, and 21 probes discriminate 1 and 10 from 8 and 9. Analysis of polymorphisms in an independent set of conserved orthologous sequences (COSs) (Choi *et al.*, 2006) confirms that the majority of the genome is not structured by population.

All probes on the Affymetrix array were mapped to the Mtr3.0 assembly of the genome using 'bowtie' (Langmead *et al.*, 2009). There were 301 055 probes that have a perfect unique hit in Mtr3.0; of these 20 208 are called as SFPs at 5% FDR, giving a polymorphism rate of 6.7%. Note that these

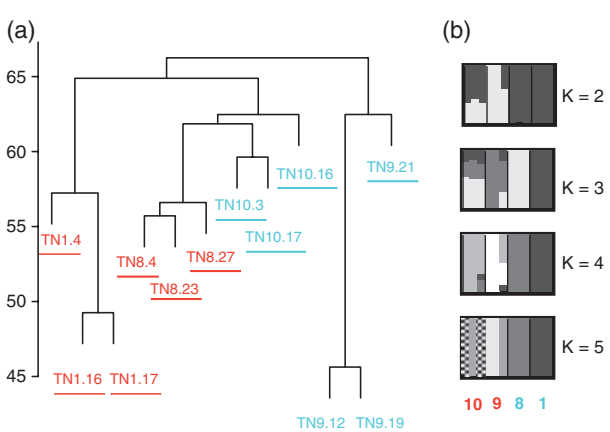

Figure 2. Population structure in 12 *Medicago truncatula* genotypes.
(a) Haplotype clustering based on the inferred markers, using the unweighted pair group method with arithmetic mean (UPGMA). Genotypes from saline sites are red and those from non-saline sites are blue.
(b) STRUCTURE analysis of a subset of the single feature polymorphisms (SFPs) showing ancestral population assignments assuming historical populations K = 2, 3, 4, and 5.

polymorphisms may include nucleotide changes, indels, and copy-number variants. To investigate patterns of LD across the genome, we compute the pairwise correlation coefficient (r^2) between SFPs across all 12 individuals. As seen in Figure 3, the maximum value is approximately 0.8 for markers within 1 kb of one another; this decays to approximately 0.4 on average by 10 kb, to approximately 0.2 by 100 kb, and to background levels by approximately 500 kb. Useful LD, i.e. correlations that would enable markers to tag causal single nucleotide polymorphisms (SNPs), thus extends on average approximately 10 kb in these populations.

Comparison between our algorithm and the *t*-test/*q*-value approach

Typically, studies of population differentiation with microarray genotyping without a reference employ *t*-tests between habitat types to determine which markers are differentiated. This test is appropriate when populations are pooled, since the *t*-test then tests the difference in allele frequencies (Turner *et al.*, 2008a,b). However, when applied to hybridization data from individuals the *t*-test confounds marker detection and habitat assortment because non-differentiated markers are not identified (e.g. Turner *et al.*, 2005); to the best of our knowledge, the statistical properties of this scenario have not been investigated. Our approach represents an alternative that yields information about overall diversity and LD across the genome in addition to habitat differentiation.

Since our AICc method is new, we compared it with the *t*-test approach by computing *t*-tests between the log-transformed hybridization intensities of individual genotypes from two saline habitats and the two non-saline habitats. After correcting for multiple testing by converting the resulting *P*-values into *q*-values, we find that overall the AIC approach calls fewer markers that assort by habitat. The intersection of the AICc and *t*-test approaches is relatively high, with the intersection containing 83 and 71% of each at a FDR of 0.05.

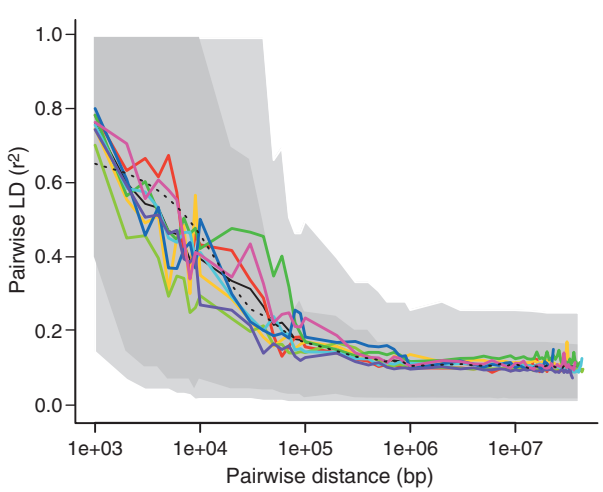

Figure 3. Pairwise correlation between markers averaged over markers within a given distance.
Each chromosome is depicted by a different solid color. Linkage is half-decayed (from 0.8 to 0.4) at approximately 10 kb. Dark gray shows the 20 to 80% quantiles across all markers, while light gray shows the 10 to 90% quantiles. The dotted line shows a second-degree lowess fit across all data with span parameter set to 0.01; linkage disequilibrium (LD) is fully decayed between markers approximately 500 kb apart.

Analysis of regions containing markers that assort by habitat

Markers that assort by habitat across multiple populations are potential evidence of selection operating on genes that confer adaptation to the corresponding habitat. Although our current data set is limited by marker density and analysis of a small set of individual genotypes, we are nevertheless able to estimate the extent of local LD and thus circumscribe a set of candidate genes for adaptation to saline habitats. A total of 52 SFPs assort with habitat and are mapped in Mt3.0; these cluster in 18 genome regions.

We define a region by the presence of one or more markers that assort by habitat, flanked by two consecutive marker probes that have low LD with the assorting focal probes ($r^2 < 0.5$). When there are habitat-assorting markers within 100 kb we consider these to be one region even if there are some low-LD markers present in the region (Table 1). These regions range in size from 229 bp (region 6.3) to 65.62 kb (region 5.1.1) with an average size of 27 kb and containing from 1 to 15 genes (average of 6.6 genes per region). LD decay around these regions is shown in Figure 4 and the complete list of 125 candidate genes is given in Table S2.

Although probe density and polymorphism rates constrain detail in the analysis, it is evident that rates of LD decay around candidate regions are reflective of our genome-wide estimates (Figure 4). Thus, correlations between the focal and linked marker probes typically decay to $r^2 < 0.3$–0.5 within 10 kb. In four regions (Figure 4, panels h, l, m and r), significant LD is apparent between probes that are separated by distances considerably greater than 10 kb (up to 100s of kb) with intervening regions of low LD. Such instances might arise from either biological (i.e. mutation or recombination) or technical (array design or sample size) circumstances, as discussed in the example below. In support of the former possibility, two of these regions (Figure 4, panels l and m) occur on chromosome 6, a genome segment that is notoriously rich in fast-evolving NBS-LRR disease resistance genes (e.g. Zhu *et al.*, 2002). NBS-LRR genes evolve by processes that can involve high rates of recombination by unequal crossing over and gene conversion, as well as diversifying selection, all of which are factors that could underlie the observed patterns of LD. Interestingly, two additional regions of chromosome 6 also include probes that assort by habitat (shown in Figure 4, panels n and o), suggesting

Table 1 Overview of the 18 genomic regions that are differentiated between saline and non-saline habitats

Region	Chr	Coordinates	Size (kb)	No. probes	No. markers	Assort with habitat	Amplicon	No. candidate genes
1.1	1	chr1:17066007..17130116	64.110	187	29	*M. truncatula*.21891.1.S1_at:635:557; *M. truncatula*.48008.1.S1_at:674:567	Chr_1	13
2.1	2	chr2:18976699..18994201	2.162	26	2	*M. truncatula*.23081.1.S1_at:1046:237; *M. truncatula*.23081.1.S1_at:421:337		5
2.2	2	chr2:21927619..21970724	43.110	68	6	*M. truncatula*.20573.1.S1_at:839:803; *M. truncatula*.20569.1.S1_at:603:565; *M. truncatula*.20569.1.S1_at:482:439	Chr_2.1; Chr2.2	6
3.1	3	chr3:21844852..21847013	2.162	19	3	*M. truncatula*.8358.1.S1_at:427:1115	Chr_3	1
3.2	3	chr3:22351489..22401448	49.960	67	2	*M. truncatula*.10504.1.S1_at:798:1029		12
4.1	4	chr4:5884997..5913860	28.860	42	10	*M. truncatula*.33441.1.S1_at:147:233		8
4.2	4	chr4:28624487..28676670	52.180	114	1	*M. truncatula*.37707.1.S1_at:1145:225		10
5.1.1	5	chr5:8617978..8683602	65.620	104	13	*M. truncatula*.48956.1.S1_at:865:623; *M. truncatula*.48956.1.S1_at:763:453; *M. truncatula*.48956.1.S1_at:1070:921; *M. truncatula*.48956.1.S1_at:464:815; *M. truncatula*.48956.1.S1_at:586:901; *M. truncatula*.48956.1.S1_at:846:457; *M. truncatula*.17919.1.S1_at:414:683	Chr_5.1.1	15
5.1.2	5	chr5:8779956..8845569	65.610	48	1	*M. truncatula*.42442.1.S1_at:633:1151	Chr_5.1.2	11
5.2	5	chr5:9272181..9296836	24.660	22	4	*M. truncatula*.43580.1.S1_at:1008:607; *M. truncatula*.43580.1.S1_at:1012:429; *M. truncatula*.43580.1.S1_at:442:351; *M. truncatula*.43580.1.S1_at:394:669		7
5.3	5	chr5:11059938..11066007	6.070	21	2	*M. truncatula*.11801.1.S1_at:95:849		2
5.4	5	chr5:30719205..30723486	4.282	11	6	*M. truncatula*.18994.1.S1_at:596:269; *M. truncatula*.18994.1.S1_at:283:687		1
6.1	6	chr6:10237963..10289680	51.720	85	23	*M. truncatula*.2455.1.S1_at:223:475; *M. truncatula*.2455.1.S1_at:369:1055; *M. truncatula*.2455.1.S1_at:14:455; *M. truncatula*.2455.1.S1_at:188:447; *M. truncatula*.2455.1.S1_at:970:139; *M. truncatula*.2455.1.S1_at:230:469; *M. truncatula*.2455.1.S1_at:71:559; *M. truncatula*.11624.1.S1_at:590:975; *M. truncatula*.11624.1.S1_at:763:435; *M. truncatula*.6977.1.S1_at:901:537; *M. truncatula*.6977.1.S1_at:97:989; *M. truncatula*.6977.1.S1_at:690:1091; *M. truncatula*.6977.1.S1_at:235:655; *M. truncatula*.6977.1.S1_at:84:1057; *M. truncatula*.6977.1.S1_at:967:359; *M. truncatula*.6977.1.S1_at:310:235; *M. truncatula*.6977.1.S1_at:727:319		10
6.2	6	chr6:10662451..10663318	0.868	10	10	*M. truncatula*.48980.1.S1_at:668:433		2
6.3	6	chr6:11332555..11332783	0.229	6	5	*M. truncatula*.22401.1.S1_at:135:569		1
6.4	6	chr6:14923810..14927210	3.401	29	11	*M. truncatula*.30229.1.S1_at:386:321		1
7.1	7	chr7:22444707..22448619	3.913	20	1	*M. truncatula*.14073.1.S1_at:689:685		2
7.2	7	chr7:26351771..26373082	21.310	84	1	*M. truncatula*.12942.1.S1_at:303:1091		9
8.1	8	chr8:868185..892217	24.030	31	9	*M. truncatula*.32725.1.S1_at:117:409; *M. truncatula*.32725.1.S1_at:678:723; *M. truncatula*.32725.1.S1_at:902:1077; *M. truncatula*.35740.1.S1_at:805:725		9

that this linkage group, which is apparently highly dynamic and not conserved in other analyzed legume genera (e.g. Choi *et al.*, 2006), may have been under selection in saline habitats. It is important to note that the saline habitats sampled here differ in many aspects from the non-saline habitats in terms of soil characteristics and vegetation composition, any of which could potentially mediate habitat-specific selection. The NBS-LRR genes mentioned above

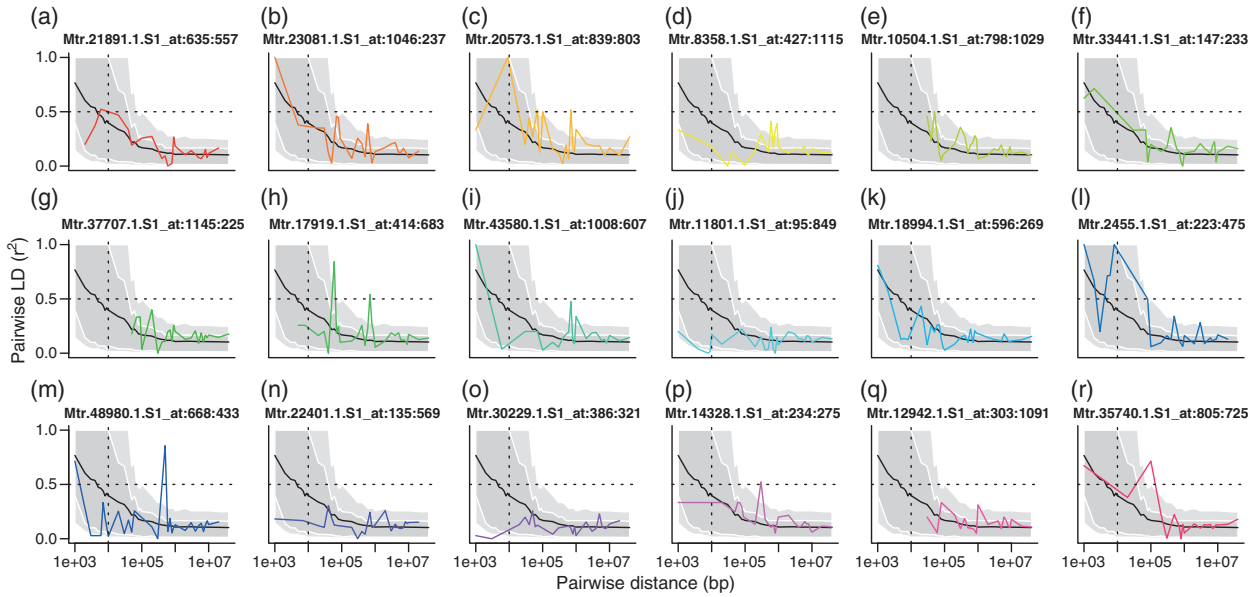

Figure 4. Linkage disequilibrium (LD) decay around focal single feature polymorphisms (SFPs) that assort by habitat.
The first habitat-assorting SFP at a locus is given in the title. The black line gives genome-wide average LD, while the gray regions are as in Figure 3 (dark gray shows the 20 to 80% quantiles across all markers, while light gray shows the 10 to 90% quantiles).

could themselves be targets of selection, but are more likely to be linked to such targets. While it is premature to view this differentiation as final evidence of selection, differentiated SFPs represent candidate genomic regions for selection based on habitat type.

As an example, we present a single 250 kb genome interval from chromosome 5 that contains eight focal marker probes that assort by habitat separated by non-assorting SFPs (Figure 5). Whether this region is actually under selection by habitat-mediated factors awaits future confir-

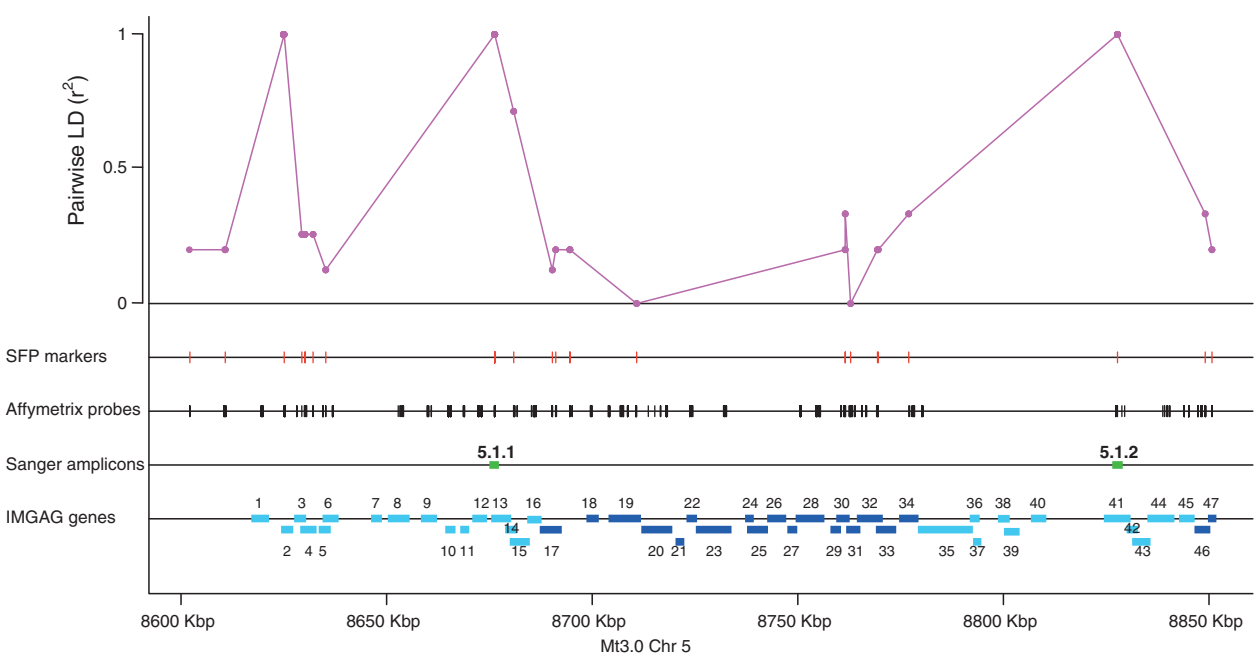

Figure 5. Regions 5.1.1 and 5.1.2 that are differentiated between saline and non-saline habitats in the four Tunisian populations.
Top: Pairwise correlation coefficient between each marker with habitat.
Bottom rows show the positions of: (i) SFPs, (ii) all uniquely mapped Affymetrix probes, (iii) the Sanger sequences regions used for validation, and (iv) International Medicago Genome Annotation Group (IMGAG) gene predictions. The IMGAG genes in cyan are considered to fall within the candidate regions, while those in dark blue are considered outside the regions. Gene numbers in this figure correspond to genes listed in Table 2.

mation, but we view it as a strong candidate and present it as an example of processes that could be acting. The left region (5.1.1) contains seven markers that assort with habitat, but these are interrupted by five marker probes in low LD with the assorting SFPs. The right-most region (5.1.2) contains a single assorting SFP, separated from region 5.1.1 by 11 SPFs in low LD with the focal SFPs.

The eight habitat-assorting SFPs detailed in Figure 5 could be the product of distinct selection events separated by historical recombination, or a single region with a common selection history and relatively unstable intervening genome features. In either case, the annotation of predicted genes obtained from the International Medicago Genome Annotation Group (IMGAG) provides a starting point for estimating gene function and narrowing the list of candidate genes in these regions for subsequent functional analysis. Table 2 lists the candidate genes from the region detailed in Figure 5. Several of the deduced proteins have potential roles in physiological and/or regulatory adjustments to abiotic stress, as well as in biotic stress responses. Ultimately it will be important to narrow the candidate genome intervals by more precise genome characterization on larger numbers of individuals, e.g. using second-generation sequencing methodologies, and to test candidate gene function by means of reverse genetic and/or biochemical characterization.

Validation of the array results by resequencing

To validate our algorithm and microarray data, and to potentially extend the correlation between habitat of origin and molecular polymorphisms, we analyzed chosen loci across a larger number of individual genotypes. Initially, Sanger resequencing was used to characterize molecular variation underlying different hybridization intensities, focusing on six genome regions (including regions 5.1.1 and 5.1.2, detailed in Figure 5) that assort between saline and non-saline populations. A total of 14 differentiated features were analyzed and polymorphisms that correlate with probe hybridization intensities were identified. The nature and location of the polymorphisms corresponding to the 12 resequenced assorting probes are described in Table 3. An additional 39 non-polymorphic probes were also confirmed in the sequencing data; all validated probes are described in Table S3. One SFP did not possess polymorphism in the Sanger data, for a false positive rate of 7%. Resequencing revealed additional polymorphism at 33% of probes; this high false negative rate is expected since probes are not sensitive to polymorphism near their edges. Of the polymorphic SFPs, one locus was shown to be tri-allelic and 7/12 genotypes at this locus were miscalled by the SFP approach. Excluding this special case, only two of the remaining 92 sequenced alleles were miscalled (2%) with 13 alleles ambiguous due to lack of PCR amplification.

To extend the correlation between genetic polymorphism and habitat of origin we analyzed a larger set of 33 individuals that were derived from the same four Tunisian populations represented by the original 12 genotypes. For each of the six genome regions, one resequenced polymorphism was selected for analysis; towards this end, SNP polymorphisms were converted to a simple allele-specific oligonucleotide assay, while insertion–deletion polymorphisms were monitored by direct resequencing. As shown in Table 4, all of the six analyzed loci revealed a significant assortment of genetic polymorphism by habitat. These results extend our initial observations, which suggest that genes contained within these regions could function in adaptation to saline habitats.

We computed standard population genetic statistics on the Sanger sequences obtained for six differentiated loci, as well as for a set of seven 'control' loci corresponding to COS markers. As shown in Table 5, some candidate loci have higher levels of polymorphism than control loci (Pi for candidate loci 0.00082–0.028, control loci, 0.0014–0.0094; Theta per nucleotide candidate loci 0.00075–0.038, control loci 0.001698–0.00857). F_{ST} tends to be higher for candidate loci than control loci (F_{ST} for candidate loci 0.32–0.71, control loci −0.03 to 0.54). However, Bonferroni-corrected *t*-tests do not support differences between candidate and control loci for any parameter, other than a marginally significant difference in F_{ST} (Pi, $P = 0.188$; ThetaNuc, $P = 0.233$; Tajima's *D*, $P = 0.515$; F_{ST}, $P = 0.0106$; *Nm*, $P = 0.310$).

CONCLUSIONS

Tunisian populations of *Medicago truncatula* harbor substantial amounts of polymorphism with relatively low levels of LD. LD is half of its maximal value at approximately 10 kb and at background levels by approximately 500 kb on average. For comparison, in global samples of the model plant *A. thaliana*, LD extends on average by approximately 10 kb (Kim *et al.*, 2007). In *Hordeum vulgare* (barley), cultivated *Hordeum* germplasm has LD extending across a 212 kb region, while in wild *Hordeum spontaneum* LD does not extend past genic regions, i.e. 28 kb (Caldwell *et al.*, 2006). Similarly, cultivated species of rice, including *Oryza indica* and tropical and temperate *Oryza japonica*, have average LDs of 75, 150, and 500 kb respectively, while their wild relative *Oryza rufipogon* has LD < 40 kb (Mather *et al.*, 2007). More limited data are available for legume species, but in the case of soybean (*Glycine max*) compared to its wild progenitor *Glycine soja*, LD extends up to 500 kb in cultivated accessions while it decays within 100 kb in non-domesticated genotypes, though different genomic regions show slightly different patterns (Hyten *et al.*, 2007). Our *M. truncatula* samples span four subpopulations, which could explain why LD takes approximately 500 kb to completely decay.

Table 2 Annotation of candidate genes within the chromosome 5 region 5.1, shown in Figure 5. Genes shaded in grey are either flanked by or border on probes that assort with saline habitats, with corresponding linkage disequilibrium (LD) values of flanking probes indicated. Gene numbers (Gene No.) correspond to numbering in Figure 5

Gene no.	Gene ID	LD	Location	IMGAG annotation
1	5g022170	0.20–1.00	8617978	Armadillo
2	5g022180	0.20–1.00	8625089	Hypothetical protein
3	5g022190	1.00–0.26	8628391	Hypothetical protein
4	5g022200	0.26–0.26	8629743	Peptidase C48
5	5g022210	0.26–1.0	8634233	Hypothetical protein
6	5g022220	1.00–1.00	8635311	FAR1
7	5g022230	1.00–1.00	8647104	HAT dimerisation
8	5g022240	1.00–1.00	8651184	Oxidoreductase FAD/NAD (P)-binding
9	5g022250	1.00–1.00	8659225	Thaumatin
10	5g022260	1.00–1.00	8665015	Hypothetical protein
11	5g022270	1.00–1.00	8668657	Short-chain dehydrogenase/reductase SDR
12	5g022280	1.00–1.00	8671722	Thaumatin
13	5g022290	1.00–0.71	8676311	Hypothetical protein
14	5g022300	1.00–0.71	8679578	Hypothetical protein
15	5g022310	1.00–0.71	8680577	Nodulin-like
16	5g022320	0.71–0.13	8684964	Rhodanese-like
17	5g022330	0.71–0.13	8687859	Mlo-related protein
18	5g022340	0.20–0.00	8699479	Cyclin-like F-box; F-box protein interaction domain
19	5g022350	0.20–0.00	8704849	von Willebrand factor
20	5g022360	0.00–0.20	8712687	Senescence-associated
21	5g022370	0.00–0.20	8720896	Hypothetical protein
22	5g022380	0.00–0.20	8723815	Hypothetical protein
23	5g022390	0.00–0.20	8725953	Response regulator receiver Heavy metal sensor kinase
24	5g022400	0.00–0.20	8738079	NB-ARC
25	5g022410	0.00–0.20	8738399	Leucine-rich repeat
26	5g022420	0.00–0.20	8743465	Heme peroxidase
27	5g022430	0.00–0.20	8748205	Hypothetical protein
28	5g022440	0.00–0.20	8750392	Protein of unknown function DUF630
29	5g022450	0.00–0.20	8758684	Hypothetical protein
30	5g022460	0.00–0.20	8760270	Peptidase C1A
31	5g022470	0.33–0.00	8762509	Hypothetical protein
32	5g022480	0.00–0.20	8765278	3-Dehydroquinate synthase
33	5g022490	0.20–0.33	8769738	D-galactoside/L-rhamnose binding SUEL lectin Glycoside hydrolase
34	5g022500	0.20–0.33	8775550	Ras GTPase
35	5g022510	0.33–1.00	8779956	Vacuolar protein sorting-associated protein 35
36	5g022520	0.33–1.00	8792564	Heavy meta transport/detoxification protein
37	5g022530	0.33–1.00	8793254	Hypothetical protein
38	5g022540	0.33–1.00	8799416	Lipolytic enzyme
39	5g022550	0.33–1.00	8800774	Lipolytic enzyme
40	5g022560	0.33–1.00	8807604	Lipolytic enzyme
41	5g022570	0.33–1.00	8825347	Uncharacterized Cys-rich domain
42	5g022580	1.00–0.33	8830696	Hypothetical protein
43	5g022590	1.00–0.33	8831951	Splicing factor motif WD40-like
44	5g022600	1.00–0.33	8835936	Diaminopimelate decarboxylase
45	5g022610	1.00–0.33	8843627	SKP1 component
46	5g022620	0.33–0.20	8847219	Cupin
47	5g022630	0.33–0.20	8850664	Hypothetical protein

IMGAG, International Medicago Genome Annotation Group.

Extending our microarray genotyping polymorphism rates to the whole genome, we predict on average 2.6 polymorphic sites kb^{-1}. High levels of linkage disequilibrium in these populations extend 10–100 kb on average, so we expect there to be 26–260 segregating sites per LD block. With a target of 10 markers per LD block and an estimated genome size of 500 Mb, a dense marker set in these *M. truncatula* populations would require half a million markers. This is readily achievable with current microarray technology or with next-generation sequencing. In addition, the observation that only 3% of our polymorphic probes assort with population suggests that gene flow among these populations is relatively high. Our data lead us to predict that genome-wide association mapping in *M. truncatula* is likely

Table 3 Sequence polymorphism associated with single feature polymorphisms (SFPs) for the six loci validated by Sanger sequencing

Region	Amplicon	Probe set	Chr	Polymorphism	Position	Saline-like	Non-saline-like	Genomic region	IMGAG annotations
1.1	Chr_1	*M. truncatula*.21891.1.S1	1	SNP	chr1:17088612	G	A	Exon/silent	Leucine-rich repeat (Medtr1g086410)
2.1	Chr_2.1	*M. truncatula*.20573.1.S1	2	Indel	chr2:21934468..21934962	Present	Absent	Intergenic	Hypothetical protein (Medtr2g095720); Calcium-binding EF-hand (Medtr2g095730)
2.1	Chr_2.2	*M. truncatula*.20569.1.S1	2	Indel	chr2:21942438..21943217	Present	Absent	Intergenic	Hypothetical proteins (Medtr2g095740 and Medtr2g095750)
3.1	Chr_3	*M. truncatula*.8358.1.S1	3	SNP	chr3:21846708	G	A	Intron	Hypothetical protein (Medtr3g092650)
5.1.1	Chr_5.1.1	*M. truncatula*.48956.1.S1	5	SNP	chr5:8676302	C	G	UTR	Hypothetical protein (Medtr5g022290)
5.1.2	Chr_5.1.2	*M. truncatula*.42442.1.S1	5	SNP	chr5:8827741	T	C	Exon/silent	Uncharacterized Cys-rich domain (Medtr5g022570)

to be successful in comprehensively localizing the genetic basis of adaptation. Indeed, our coarse survey has already yielded several plausible candidates for local adaptation to soil salinity.

EXPERIMENTAL PROCEDURES

Plant genotypes

Genotypes were collected in Tunisia in July 1999 by MEA with the assistance of Chedly Abdelly. At each of 10 sites, 30–100 pods of *M. truncatula* were collected at random in an area of radius 500 m. Pods were selected to maximize the variation in pod morphology at a site, thus minimizing the chance that pods from the same parent are sampled. Twelve lines per site were multiplied in the greenhouse; although *M. truncatula* is highly selfing in nature, each line was selfed twice to lower remaining heterozygosity. The collection is housed at the CBBC (Centre of Biotechnology of Borj Cedria, Tunisia) and germplasm is available upon request (contact Dr Mounawer Badri). From among the 10 sites, four populations were selected, each in the north of Tunisia: TN1 (Enfidha), TN8 (Soliman), TN10 (Rhayet), and TN9 (Bulla Regia). Enfidha and Soliman sites have highly saline soil (8.65 and 4.40 g l^{-1}) while Rhayet and Bulla Regia sites have low levels of salt (0.80 and 0.95 g l^{-1} (Lazrek *et al.*, 2009) (Figure 1).

Microarray experiment

Genomic DNA was extracted from young leaves grown in growth rooms (ENSA Toulouse, France) using the DNeasy Plant Mini Kit (Qiagen, http://www.qiagen.com/). Genomic DNA was subsequently amplified using the Repli-g Midi (Qiagen). Amplified DNA samples from individual genotypes were extracted with phenol–chloroform, and the resulting purified DNA was fragmented by partial digestion with DNase, as follows: 10.5 μg DNA was dissolved in 30 μl double-distilled (dd) H_2O, plus 4 μl One-Phor-All buffer (Amersham Biosciences 27-0910-02, http://www.gelifesciences.com), 0.2975 μl DNase (Promega M6101-RQ1, http://www.promega.com/), and 0.14 μl acetylated BSA (Invitrogen 15561-020, http://www.invitrogen.com/). The DNase digestion was allowed to proceed for 16 min at 37°C, followed by heat inactivation at 99°C for 15 min, and cooling to 12°C for 15 min. All reactions were carried out in a MJ Research thermocycler (Waltham, MD, USA). Three microliters of each DNA sample was visualized by ethidium bromide staining, following separation by gel electrophoresis on a 4% agarose SFR 0.5 TRIS-borate-EDTA (TBE; TRIS = 2-amino-2-(hydroxymethyl)-1,3-propanediol) gel (50 V for 120 min) with 10 bp and 100 bp DNA ladders (Promega). Samples that yielded bright smears from 20 to 100bp were selected and labeled with biotin by adding 2μl Biotin-N6-ddATP (Enzo 42809) and 3 μl RTdT (diluted to 15U μl^{-1}; Promega M1875) and running the following program in a MJ Research thermocycler: 90 min at 37°C, 15 min at 99°C, 5 min at 12°C. Labeled samples were frozen and delivered on dry ice to the Microarray Core facility at Children's Hospital, Los Angeles, USA, where they were hybridized to Medicago Genechips (Affymetrix) using the Affymetrix Hyb, Wash, and Stain Kit with the following hybridization cocktail: 125 μl 2 × hybridization mix, 4.17 μl control oligo B2, 12.5 μl 20 × hybridization controls, 25 μl DMSO, labeled target DNA (9.585 μg), ddH2O to 250 μl, and wash protocol FS450_0001. A single array was hybridized with each individual genotype's DNA to maximize the number of individuals assayed for a given cost. While this does not enable the estimation of technical error in genotype calls, the resulting data are sufficient to identify many SFPs that are replicated at the population level, namely those that occur in two or more individuals.

Validation of the SFPs

Genomic DNA was extracted from 45 individuals from four Tunisian populations using the DNeasy Plant Mini kit (Qiagen). Six loci around Affymetrix probes and seven COS markers were amplified in the genotypes used for the microarray experiment. Primer3Plus software was used to design primers to amplify 700–1000 bp around the Affy probes *M. truncatula*.21891.1.S1 (F, tatcagagga agctgcaaaagc; R, tcagcctcttcatcaatgtcc), *M. truncatula*.48956.1.S1 (F, ttgacagctacaacaaggaagc; R, gtaacctttctcccaaagttgc), *M. truncatula*.42442.1.S1 (F, ctcttccggacaagtgttcacc; R, cacaagccacaaacacataagagc), *M. truncatula*.20573.1.S1 (F, tctctactagttccctctctattagttcc; R, cagtaaaaatcgcgctacgg), *M. truncatula*.20569.1.S1 (F, tctgccatagcc atgtttcg; R, aaccggtcatcttacacaacg), and *M. truncatula*.8358.1.S1 (F, taaacccatcagtcccatcacc; R, tgtagatttgttgttggcaagg). The COS markers 1433P, AAT, AGT, CALTL, CNGC4, SHMT, and SUSY were selected

Table 4 Expanded genotyping of polymorphisms that assort with habitat. Sites were identified by Sanger sequencing of loci containing the probe sets that assort with habitat in 45 TN genotypes from saline (TN1 and TN8) and non-saline (TN9 and TN10) populations, from which 12 (in bold) were used in the microarray experiment

Probe set	Chr.	*P*-value	**TN 1.4**	**TN 1.6**	**TN 1.7**	TN 1.3	TN 1.5	TN 1.8	TN 1.9	TN 1.11	TN 1.12	TN 1.14	TN 1.19	TN 1.20	TN 1.21	**TN 8.4**	**TN 8.23**	**TN 8.27**	TN 8.3	TN 8.15	TN 8.20	TN 8.21	TN 8.22	TN 8.24	TN 8.29	**TN 9.12**	**TN 9.19**	**TN 9.21**	TN 9.1	TN 9.3	TN 9.4	TN 9.5	TN 9.17	TN 9.18	TN 9.20	TN 9.22	**TN 10.3**	**TN 10.16**	**TN 10.17**	TN 10.1	TN 10.2	TN 10.4	TN 10.8	TN 10.9	TN 10.15	TN 10.19	TN 10.23
Mtr.21891.1.S1	1	1.71×10^{-2}	S	S	S	S	NS	NS	S	S	S	S	NS	NS	NS	S	S	S	NS	NS	S	NS	S	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Mtr.20573.1.S1	2	3.70×10^{-5}	S	S	S	NS	NS	NS	S	S	S	S	NS	S	S	S	S	S	S	NS	S	NS	S	NS	S	NS	NS	NS	S	NS	S	S	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Mtr.20569.1.S1	2	1.10×10^{-7}	S	S	S	NS	S	S	S	S	S	S	S	S	S	S	S	S	S	NS	S	NS	S	S	NS	NS	NS	NS	S	NS	S	S	NS	NS	NS	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Mtr.8358.1.S1	3	5.37×10^{-7}	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	–	S	S	–	S	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Mtr.48956.1.S1	5	2.18×10^{-3}	S	S	S	S	NS	NS	S	S	–	–	–	S	–	S	S	S	–	–	S	–	S	–	S	NS	NS	NS	NS	–	S	NS	S	NS	NS	NS	NS	NS	NS	NS	NS	–	NS	NS	NS	NS	NS	NS
Mtr.42442.1.S1	5	5.18×10^{-7}	S	S	S	S	S	S	S	S	S	S	S	S	NS	S	S	S	S	S	S	NS	S	S	S	NS	NS	NS	S	NS	S	NS	S	S	NS	S	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

S, saline-like; NS, non-saline like; –, putative heterozygote; grey, sequences confirmed by Sanger sequencing; *P*-values from Fisher's exact test, Bonferroni corrected, on data from the 33 individuals not used on the arrays.

(Choi *et al.*, 2006). The PCR reactions were performed in a Tetrad 2 Thermal Cycler PTC-0240G using Takara's (http://www.takara-bio.com/) Ex Taq® DNA Polymerase (3 min at 95°C; 40 cycles of 30 sec 95°C, 30 sec 55°C (except for *M. truncatula*.20573.1.S1 and *M. truncatula*. 20569.1.S1 where 57°C and 60°C were used, respectively), 90 sec 68°C, and 3 min 68°C. Amplicons (3 μl) were visualized by gel electrophoresis on a 1.2% agarose 0.5 TBE gel running at 100 V for 30–40 min with All Purpose Hi-Lo® DNA Marker (Bionexus, http://www.bionexus.net/). The PCR products were cleaned using 0.5 U shrimp alkaline phosphatase (SAP) (USB, http://www.usb-web.com/), SAP buffer (USB), and 0.2 U Exo I (USB) (30 min, 37°C), and were sequenced using an ABI 3730 XL capillary sequencer (Applied Biosystems, http://www.appliedbiosystems.com/). Sequences were analyzed using CodonCode Aligner 2.0.6 and mapped into Mt3.0 using BLAST. Genes around the sequenced Affymetrix probes were assessed using the IMGAG Genome Annotation Version 3.0.

The identified SNPs were surveyed in the 45 individuals from the four TN populations using the ABI Prism® SNaPshot® Multiplex kit: 1 μl SNaPshot mix, 2 μl of clean PCR product and 2 pmol of extension primers for *M. truncatula*.21891.1.S1 (TTTGAAGGAATCTGCACC), *M. truncatula*.48956.1.S1 (GTTGGACGTGGTGGCGAGCTTA), *M. truncatula*.42442.1.S1 (TCTACTTGCTTGTTGTTC), *M. truncatula*.20573.1.S1 (AAAATGCAACTGGAAATAAGAC), and *M. truncatula*.8358.1.S1 (TCTTTCACTATTACTTCAACTA). The extension reaction consisted of 40 cycles of 10 sec 95°C, 5 sec 50°C, and 30 sec 60°C. The individuals were genotyped using an ABI 3730 XL capillary sequencer (Applied Biosystems) and the polymorphisms were analyzed in the GeneMapper® software v.3.7 (Applied Biosystems). Fisher's exact test Bonferroni corrected on the 33 independent inbred lines was performed to assess the statistical significance of the distribution of the polymorphisms with respect to habitat type.

Microarray data analysis

We first examined the .jpeg for each array for defects and found none. All statistics were performed in R 2.6.2 (R Team 2009); code is available upon request. The Affymetrix® GeneChip Medicago Genome Array contains 673 880 probe pairs in total; 560 206 of these (50 902 probe sets) are specific to *M. truncatula* with an additional 1896 probe sets specific to *M. sativa*, 8305 probe sets specific to the bacterial symbiont *Sinorhizobium meliloti*, and 14 control probe sets. A comparison of raw perfect-match (pm) probe intensities and mismatch (mm) probe intensities revealed that 0.8608 of targets hybridized more strongly to the pm probe; when only *M. truncatula* specific probe pairs were considered the pm intensity was greater than the mm intensity for 0.9148 of targets, indicating a substantial amount of signal in our data. Examining the log pm intensity distribution did not reveal large differences between arrays. We used two standard background corrections implemented in Bioconductor (Gentleman *et al.*, 2004), rma and mas. Rma ('robust multi-array averaging') models each pm intensity as having signal and error components, while mas (Affymetrix's 'Micro Array Suite') performs a spatial correction for each array by considering the lowest-intensity probes in each grid. We performed two standard normalizations: slide mean normalization, to scale each slide to have the same mean intensity, and quantile normalization, which scales each slide to have the same intensity distribution. The six potential combinations (raw, rma, mas correction by slide mean, quantile normalization) were compared. Mas correction changed the second peak corresponding to low hybridization intensities into a shoulder that obscured differences between strong and weak hybridization signals, so it was not used in further analyses. Raw and rma histograms had similar shapes to one another, with clear peaks for normal

Table 5 Population genetic parameters for sequences covering differentiated single feature polymorphisms (SFPs) and control genomic regions (conserved orthologous sequence (COS) markers)

Locus	*N*_i	*N*_pops	Sites	NetSites	*S*	Pi	ThetaNuc	Tajima *D*	F_{ST}	*Nm*
Control loci										
COS1_1433	12	4	480	480	10	0.003756	0.006898	−1.874	−0.02061	−12.38
COS10_SHMT	13	4	565	564	15	0.00941	0.00857	0.4096	0.14493	1.48
COS11_SUSY	12	4	490	487	4	0.001368	0.002719	−1.7469	*n.d.*	*n.d.*
COS2_AAT	12	4	807	807	7	0.002008	0.002872	−1.1763	−0.03448	−7.5
COS3_AGT	12	4	823	724	5	0.002218	0.002286	−0.1105	0.07954	2.89
COS6_CALTL	12	4	496	390	2	0.00202	0.001698	0.5542	0.18182	1.13
COS7_CNGC4	11	4	313	303	2	0.00228	0.002253	0.0361	0.53846	0.21
Differentiated loci										
M. truncatula.20569.1.S1	21	3	741	741	2	0.000822	0.00075	0.2222	0.71247	0.1
M. truncatula.20573.1.S1	19	3	479	477	31	0.023269	0.018594	0.9991	0.3169	0.54
M. truncatula.21891.1.S1	24	4	625	625	4	0.001472	0.001713	−0.3852	0.48108	0.27
M. truncatula.42442.1.S1	24	4	583	582	6	0.001556	0.00276	−1.3194	0.43333	0.33
M. truncatula.48956.1.S1	23	4	501	500	8	0.004695	0.004335	0.2698	0.3704	0.42
M. truncatula.8358.1.S1	24	4	383	383	50	0.028096	0.037756	−0.9958	0.50345	0.25

*N*_i, number of individuals with Sanger sequence data; *N*_pops, number of populations, Sites, total length of sequenced locus; NetSites, sites with no missing data; *S*, number of segregating sites; Pi, average pairwise nucleotide polymorphism; ThetaNuc, per nucleotide estimate of Watterson's theta; Tajima *D*, measure of allele frequency skew from neutral; F_{ST}, measure of population differentiation; *Nm*, estimate of gene flow between populations (*m*) scaled by effective population size (*N*), n.d., not determined.

hybridization intensities and weak intensities that presumably correspond to sequence divergence.

For each data processing, more markers are called for quantile normalization than for slide-mean normalization with many of these markers present in only two of the 12 genotypes. Since low-frequency markers will deflate estimates of LD, we focus on the slide-mean normalized raw data for the analyses in this paper.

Algorithm for determination of SFPs

We develop a new algorithm to determine whether the 12 individuals are polymorphic at a site as reflected by Affymetrix probe hybridization intensity. We presume that each individual is homozygous, since *M. truncatula* is highly selfing in nature and wild plant genotypes were further selfed for at least two generations in the greenhouse. Our method uses simple information criteria to compare two types of models: Model_1 where a probe does not cover a polymorphism and Model_2 where a probe does detect polymorphism and is therefore a marker. Information criteria are used in model selection to balance the explanatory power of each model against the number of model parameters, thus identifying which model is closer to the truth. Since this is a non-parametric procedure without established significance cutoffs, we use simulation to determine the significance threshold.

For each probe, we first order the log-transformed hybridization intensities $\boldsymbol{I}$ from lowest to highest (I_1, I_2, …, I_{12}). We then consider all possible two-way splits that divide the data into contiguous sets of values, where each subset contains at least two observations. For 12 observations there are 10 possible splits [(I_1, I_2),(I_3, …, I_{12})]; [(I_1, I_2, I_3),(I_4,…, I_{12})]; …; [(I_1, …, I_{10}), (I_{11}, I_{12})]. We next calculate the likelihood of the data under Model_1, where the data are drawn from a single Normal (μ, θ) model, where μ = mean ($\boldsymbol{I}$) and θ = standard deviation ($\boldsymbol{I}$). The likelihood is then the product of probabilities:

$$L(I|\text{Model}_1) = \prod_{k=1}^{12} \Pr(I_k|\text{Normal}(\mu,\theta))$$

For each of the ten two-group models, Model_{2_i} with i = 1, …, 10, we assume that the data are drawn from two groups. For example, under Model_{2_1}, group g_1 is $\boldsymbol{I}_{g1}$ = (I_1, I_2) and group g_2 is $\boldsymbol{I}_{g2}$ = (I_3, …, I_{12}). Each group g_i is assumed to have a Normal (μ_{gi},θ_{gi}) distribution, where μ_{gi} = mean ($\boldsymbol{I}_{gi}$) and θ_{gi} = standard deviation ($\boldsymbol{I}_{gi}$). The likelihood of the data under this model is

$$L(I|\text{Model}_{2\_1}) = \prod_{k=1}^{2} \Pr(I_k|\text{Normal}(\mu_{g1},\theta_{g1})) \prod_{k=3}^{12} \Pr(I_k|\text{Normal}(\mu_{g2},\theta_{g2}))$$

Next, we use Akaike's AIC with the small sample bias correction term (Burnham and Anderson, 2002):

$$\text{AICc} = -2\text{Ln}(\text{Likelihood}(\text{Data}|\text{Model})) + 2k + 2k(k+1)/(n-k-1)$$

where k is the number of parameters in Model and n is the number of observations. We assign the model with the lowest AICc value to have score of 0 and subtract this AICc value from the AICc value for the other models to obtain $\text{delta}_{\text{AICc}}$ values. Models with higher AICc values and thus larger $\text{delta}_{\text{AICc}}$ values should be rejected; we use simulation to determine the significance threshold of $\text{delta}_{\text{AICc}}$.

To formulate a null distribution for the $\text{delta}_{\text{AICc}}$ test statistic, we simulate extensively under the no-polymorphism model, i.e. a single normal distribution having a mean and standard deviation drawn from the empirical distribution of means and standard deviations. This ensures that the range of variances in the simulation reflects the variances present in the data set. We simulated 5 million data sets and calculated the $\text{delta}_{\text{AICc}}$ values for Model_1, Model_{2_1}, …, Model_{2_10}. In order to set the significance threshold, we employ false discovery rate (FDR) criteria. The FDR is the frequency of false rejections of the null hypothesis within all null hypothesis rejections. We calculate it by dividing the expected frequency of false positives, as seen in the simulated null distribution, by the number of probes in the real data that exceed each threshold and then selecting the threshold giving the desired FDR. For example, setting the $\text{delta}_{\text{AICc}}$ threshold to 28.0812 gives on average 1568 (out of 560 206) simulated null model probes that have higher $\text{delta}_{\text{AICc}}$ values while 31 419 of the 560 206 empirical probes have $\text{delta}_{\text{AICc}} \geq 28.0812$. Since 1568/31 419 = 0.05, this $\text{delta}_{\text{AICc}}$

threshold corresponds to 5% FDR. Affymetrix probes whose hybridization data support a two-group model with delta$_{AICc} \geq 28.0812$ are referred to as 'markers' or equivalently 'SFPs'.

Analysis of population genetics and structure

Population genetic parameters were calculated using DNAsp v. 5.0 (Librado and Rozas, 2009) running under 'wine' on a Mac OSX PC. Only polymorphic sites supported by both strands of Sanger sequencing were included. Results are presented in Table 5. To explore population structure, the Bayesian clustering program STRUCTURE v. 2.3 (Pritchard *et al.*, 2000) was compiled and run on a 64-bit node. We ran each number of clusters (K) three times. As STRUCTURE has a maximum marker number of 10 000 and further requires that markers be unlinked, we sampled one marker per 50 kb (genome-wide LD is approximately 0.2 at this distance, see Figure 3) for a total of 3429 markers. For K = 1, 2, 3, 4, 5, and 6 the ln probabilities of the data were as follows: run1, −48445.5, −44279.7, −41359.8, −39861.5, −37868.6, −37333.1; run2, −48428, −44341.3, −41357.8, −39438.7, −37759.5, −37070.1; run3, −48435.2, −44353.4, −41305.3, −39497.5, −37787.1, −37744.5. As these numbers are congruent between runs, we conclude that the model converged. The 'optimal' number of clusters is five, since the change in probability between four and five is much greater than between five and six (Pritchard *et al.*, 2010). However, since 12 is a small number of individuals relative to the number of groups tested we do not have a high degree of confidence in these clusters. Individual assignment was broadly consistent between runs; representative assignments plotted with 'distruct' (Rosenberg, 2004) are shown in Figure 2b.

ACKNOWLEDGEMENTS

MLF acknowledges funding from NSERC PGSD and the UC Davis Center for Population Biology. MC acknowledges Fundação para a Ciência e Tecnologia SFRH/BD/39905/2007. This work was supported by National Science Foundation Plant Genome Research Program No. 0820846 and National Science Foundation Office of International Science and Engineering No. 0751073 awards to SVN. We would also like to thank the following people: Carine Ameline-Torregrosa (ENSAT) helped with DNA extractions; Thomas L. Turner (UCSB) provided the experimental protocol and provided advice; Matt Rolston (UCD) provided the hybridization recipe and protocol; Betty Schaub (CHLA) performed the hybridizations; Yuri Bendana (USC) implemented a local instance of *M. truncatula* GBrowse.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table S1. All probes determined to be markers in our experiment.

Table S2. All genes, probes, and markers contained in candidate regions, i.e. regions defined by markers that assort with saline source habitat.

Table S3. All probes whose marker state was verified by Sanger resequencing.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

AUTHOR CONTRIBUTIONS

MLF, SVN, MB, TH, and MEA conceived the study. TH, MB, and MEA provided material. MLF performed the microarray experiment and analyzed the data with input from SVN. MC validated the array data and gathered additional sequence data with input from RVP and DRC. MLF, MC, DRC, and SVN wrote the paper with input from all authors.

REFERENCES

Aranzana, M.J., Kim, S., Zhao, K. *et al.* (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* **1**, e60.

Badri, M., Ilahi, H., Huguet, T. and Aouani, M.E. (2007) Quantitative and molecular genetic variation in sympatric populations of Medicago laciniata and M. truncatula (Fabaceae): relationships with eco-geographical factors. *Genet. Res.* **89**, 107–122.

Bonnin, I., Huguet, T., Gherardi, M., Prosperi, J.M. and 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.

Bonnin, I., Ronfort, J., Wozniak, F. and Olivieri, I. (2001) Spatial effects and rare outcrossing events in Medicago truncatula (Fabaceae). *Mol. Ecol.* **10**, 1371–1383.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* **13**, 513–523.

Borevitz, J.O., Hazen, S.P., Michael, T.P. *et al.* (2007) Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA*, **104**, 12057–12062.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*. New York: Springer.

Caldwell, K.S., Russell, J., Langridge, P. and Powell, W. (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, Hordeum vulgare. *Genetics*, **172**, 557–567.

Chaulet, E. and Prosperi, J.M. (1994) Genetic diversity of a collection of Medicago trunculata Gaernt from Algeria. *EUCARPIA*, Evaluation and exploitation of genetic resources pre-breeding, Clermont Ferrand (France), 15–18 Mar 1994.

Chen, C., Fan, C., Gao, M. and Zhu, H. (2009) Antiquity and function of CASTOR and POLLUX, the twin ion channel-encoding genes key to the evolution of root symbioses in plants. *Plant Physiol.* **149**, 306–317.

Choi, H.K., Luckow, M.A., Doyle, J. and Cook, D.R. (2006) Development of nuclear gene-derived molecular markers linked to legume genetic maps. *Mol. Genet. Genomics*, **276**, 56–70.

De Mita, S., Santoni, S., Hochu, I., Ronfort, J. and Bataillon, T. (2006) Molecular evolution and positive selection of the symbiotic gene NORK in Medicago truncatula. *J. Mol. Evol.* **62**, 234–244.

Gentleman, R., Carey, V., Bates, D. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

Gomez, K.S., Javot, H., Deewatthanawong, P., Torres-Jerez, I., Tang, Y., Blancaflor, E.B., Udvardi, M.K. and Harrison, M.J. (2009) Medicago truncatula and Glomus intraradices gene expression in cortical cells harboring arbuscules in the arbuscular mycorrhizal symbiosis. *BMC Plant Biol.* **9**, 10.

Heath, K.D. and Tiffin, P. (2007) Context dependence in the coevolution of plant and rhizobial mutualists. *Proc. R. Soc. Lond. B Biol. Sci.* **274**, 1905–1912.

Heath, K.D. and Tiffin, P. (2009) Stabilizing mechanisms in a legume-rhizobium mutualism. *Evolution*, **63**, 652–662.

Hyten, D., Choi, I., Song, Q., Shoemaker, R., Nelson, R., Costa, J., Specht, J. and Cregan, P.B. (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics*, **175**, 1937–1944.

Johnson, N.C., Wilson, G.W., Bowker, M.A., Wilson, J.A. and Miller, R.M. (2010) Resource limitation is a driver of local adaptation in mycorrhizal symbioses. *Proc. Natl. Acad. Sci. USA*, **107**, 2093–2098.

Kim, S., Plagnol, V., Hu, T.T., Toomajian, C., Clark, R.M., Ossowski, S., Ecker, J.R., Weigel, D. and Nordborg, M. (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.* **39**, 1151–1155.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Lazrek, F., Roussel, V., Ronfort, J., Cardinet, G., Chardon, F., Aouani, M.E. and Huguet, T. (2009) The use of neutral and non-neutral SSRs to analyse the genetic structure of a Tunisian collection of Medicago truncatula lines and

to reveal associations with eco-environmental variables. *Genetica*, **135**, 391–402.

Lesins, K. and Lesins, I. (1979) *Genus Medicago (Leguminosae), A Taxogenetic Study*. The Hague, The Netherlands: W. Junk.

Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Mather, K., Caicedo, A., Polato, N., Olsen, K., McCouch, S. and Purugganan, M.D. (2007) The extent of linkage disequilibrium in rice (Oryza sativa L.). *Genetics*, **177**, 2223–2232.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pritchard, J.K., Wen, X. and Falush, D. (2010) *Documentation for Structure Software: Version 2.3*. Chicago, IL.

R Team (2009) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rangin, C., Brunel, B., Cleyet-Marel, J., Perrineau, M. and Bena, G. (2008) Effects of Medicago truncatula genetic diversity, rhizobial competition, and strain effectiveness on the diversity of a natural Sinorhizobium species community. *Appl. Environ. Microbiol.* **74**, 5653–5661.

Ronfort, J., Bataillon, T., Santoni, S., Delalande, M., David, J.L. and Prosperi, J. (2006) Microsatellite diversity and broad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in Medicago truncatula. *BMC Plant Biol.* **6**, 28.

Rosenberg, N.A. (2004) distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes*, **4**, 137–138.

Rosenberg, N.A. and Nordborg, M. (2006) A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics*, **173**, 1665–1678.

Siol, M., Bonnin, I., Olivieri, I., Prosperi, J.M. and Ronfort, J. (2007) Effective population size associated with self-fertilization: lessons from temporal changes in allele frequencies in the selfing annual Medicago truncatula. *J. Evol. Biol.* **20**, 2349–2360.

Siol, M., Prosperi, J.M., Bonnin, I. and Ronfort, J. (2008) How multilocus genotypic pattern helps to understand the history of selfing populations: a case study in Medicago truncatula. *Heredity*, **100**, 517–525.

Slatkin, M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat. Rev. Gen.* **9**, 477–485.

Small, E. and Jomphe, M. (1989) A synopsis of the genus Medicago (Leguminosae). *Can. J. Bot.* **67**, 3260–3294.

Tester, M. and Davenport, R. (2003) Na+ tolerance and Na+ transport in higher plants. *Ann. Bot.* **91**, 503–527.

Turner, T.L., Hahn, M.W. and Nuzhdin, S. (2005) Genomic islands of speciation in Anopheles gambiae. *PLoS Biol.* **3**, e285.

Turner, T.L., Levine, M.T., Eckert, M.L. and Begun, D.J. (2008a) Genomic analysis of adaptive differentiation in Drosophila melanogaster. *Genetics*, **179**, 455–473.

Turner, T.L., von Wettberg, E.J. and Nuzhdin, S. (2008b) Genomic analysis of differentiation between soil types reveals candidate genes for local adaptation in Arabidopsis lyrata. *PLoS ONE*, **3**, e3183.

Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T. and Nuzhdin, S.V. (2010) Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat. Genet.* **42**, 260–263.

Wilczek, A.M., Roe, J.L., Knapp, M.C. *et al.* (2009) Effects of genetic perturbation on seasonal life history plasticity. *Science*, **323**, 930–934.

Winzeler, E. (1998) Direct Allelic Variation Scanning of the Yeast Genome. *Science*, **281**, 1194–1197.

Young, N.D. and Udvardi, M. (2009) Translating Medicago truncatula genomics to crop legumes. *Curr. Opin. Plant Biol.* **12**, 193–201.

Zhao, K., Aranzana, M.J., Kim, S. *et al.* (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* **3**, e4.

Zhu, H., Cannon, S.B., Young, N.D. and Cook, D.R. (2002) Phylogeny and genomic organization of the TIR and non-tIR NBS-LRR resistance gene family in Medicago truncatula. *Mol. Plant Microbe Interact.* **15**, 529–539.

Accession numbers GenBank: GU133069-GU133204, ArrayExpress: E-MEXP-2686.

