Genomic signature of adaptation to climate in *Medicago truncatula*

Jeremy B. Yoder,* John Stanton-Geddes,§ Peng Zhou,† Roman Briskine,‡ Nevin D. Young,* †

Peter Tiffin*

* Department of Plant Biology, University of Minnesota, Saint Paul, MN 55108, USA
§ Department of Biology, University of Vermont, Burlington, VT 05405, USA
† Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108, USA
‡ Department of Computer Science and Engineering, University of Minnesota, Saint Paul, MN 55455, USA
Competing interests: The authors declare that no competing interests exist.

Funding: The work was funded by National Science Foundation Grants 082005 and 1237993. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

8,283 words, 4 figures, 2 tables, and 14 supplementary items

Running head: Genomics of adaptation to climate

Key words: genome scan, landscape genomics, association genetics, balancing selection, environmental gradient

Correspondence should be directed to

Jeremy B. Yoder
jbyoder@umn.edu

or

Peter Tiffin
ptiffin@umn.edu

University of Minnesota
Department of Plant Biology
1445 Gortner Avenue
Suite 250
Saint Paul, Minnesota, 55108
USA
ABSTRACT

Local adaptation and adaptive clines are pervasive in natural plant populations, yet the effects of these types of adaptation on genomic diversity are not well understood. With a dataset of 202 accessions of *Medicago truncatula* genotyped at almost 2 million single-nucleotide polymorphisms, we used mixed linear models to identify candidate loci responsible for adaptation to three climatic gradients (annual mean temperature, AMT; precipitation in the wettest month, PWM; and isothermality, ITH) representing the major axes of climate variation across the species’ range. Loci with the strongest association to these climate gradients tagged genome regions with high sequence similarity to genes with functional roles in thermal tolerance, drought tolerance, or resistance to herbivores of pathogens. Genotypes at these candidate loci also predicted the performance of an independent sample of plant accessions grown in climate-controlled conditions. Compared to a genome-wide sample of randomly drawn reference SNPs, candidates for two climate gradients, AMT and PWM, were significantly enriched for genic regions; and genome segments flanking genic AMT and PWM candidates harbored less nucleotide diversity, elevated differentiation between haplotypes carrying alternate alleles, and an overrepresentation of the most common haplotypes. These patterns of diversity are consistent with a history of soft selective sweeps acting on loci underlying adaptation to climate, but not with a history of long-term balancing selection.
INTRODUCTION

Local and clinal adaptation is widespread in natural populations (Clausen et al. 1941; Leimu and Fischer 2008) and by definition it results from selection that varies across a species’ range. Most methods to search for the targets of adaptation are designed to identify gene regions that have experienced “hard” selective sweeps, in which selection acts on new mutations that confer a selective advantage across the entire range of a sample (Maynard Smith and Haigh 1974; Nielsen 2005; Pritchard and Di Rienzo 2010; Kelly et al. 2013). These methods are not designed to identify targets of adaptation to selective environments that vary across the range of sampled populations. The targets of locally variable selection may either be maintained as stable polymorphisms or experience partial, or “soft,” sweeps either because local adaptation involves fixation of different alleles in different portions of a species’ range or because selection acts on standing variation (Hermisson and Pennings 2005; Pavlidis et al. 2012; Messer and Petrov 2013).

Identifying the molecular targets of clinal adaptation offers an opportunity to not only identify functionally important genes, but also to further our understanding of adaptation itself. If the selective environment that drives clinal adaptation is stable, and alleles responsible for adaptation are at stable equilibria, then the loci responsible for adaptation may bear population-genetic signatures of balancing selection: elevated differentiation between haplotypes linked to the alternative alleles at each causal locus (Tian et al. 2002; Hedrick 2006; Charlesworth et al. 2012). Alternatively, if the selective optimum changes over time, alleles at loci that contribute to adaptation are more likely to experience repeated partial sweeps, increasing the frequency of different adaptive alleles in different parts of the range. These partial, local sweeps may create elevated differentiation between allelic haplotypes, as in the case of balancing selection, but they
are also expected to reduce nucleotide and haplotype diversity, and create extended linkage in the genomic regions flanking each selected locus (Pennings and Hermisson 2006; Garud et al. 2013).

The increasing availability of population-genomic data has prompted the development of analytical methods to identify putative targets of local and clinal adaptation (reviewed by De Mita et al. 2013; Jones et al. 2013). Most methods to “scan” genomes for targets of clinal adaptation seek to identify adaptive loci by testing for strong association of allele frequencies with environmental variables (Frichot et al. 2013; Günther and Coop 2013) or excess population differentiation as in F_{ST} outlier approaches (Lewontin and Krakauer 1973; Beaumont 2005; Foll and Gaggiotti 2008; Excoffier et al. 2009). These approaches require that sampled individuals be divided into discrete populations, based either on prior information about sampling locations or population structure revealed in the data, and their results can be sensitive to how individuals are assigned to subpopulations (Yang et al. 2012). One way around this difficulty is to take samples from discrete habitats or the extremes of an environmental gradient (e.g., Rieseberg et al. 1999; Hohenlohe et al. 2010; Chen et al. 2012; Kujala and Savolainen 2012). While a focus on discrete environments may help to identify functionally important genes it can miss selectively important genes mediating adaptation to other environmental conditions and it may not provide insight into how selection shapes genetic diversity across environmental clines.

For samples that are distributed continuously across the landscape rather than collected from discrete subpopulations, it may be possible to identify the loci responsible for local adaptation as those displaying unusually abrupt changes in allele frequency across geographic location (Yang et al. 2012) or to search for associations between molecular variants and environmental variables using a linear modeling framework (Joost et al. 2007; Eckert et al. 2010; Yang et al. 2012; Frichot et al. 2013). Although such an approach does not specifically model
isolation by distance (see discussion in Günther and Coop 2012), mixed linear models are able to control for the potentially confounding effects of common demographic history by including a summary of among-individual relatedness (i.e., kinship or population structure) as a covariate (Zhang et al. 2010). Linear-model based scans for targets of local or clinal adaptation are also computationally efficient for very large data sets (Yang et al. 2012).

Simulations conducted by DeMita et al. (2013) suggests that in some circumstances linear models may be more powerful for identifying targets of clinal selection than approaches that search for elevated differentiation among subpopulations. In comparison to differentiation-based methods, a logistic regression approach testing for genotype-environment associations (Joost et al. 2007) had greater power to identify selected loci when the environmental selection gradient was weak. This approach performed particularly well with selfing species, and when only each sampling location was represented by a single individual (De Mita et al. 2013). However, both De Mita et al. (2013) and a similar analysis by Meirmans (2012) found that linear model methods have slightly higher risks of false positives in the presence of population structure. This suggests that a linear modeling method to test for genotype-environment associations that also controls for effects of population structure may be the most effective approach to identify targets of selection exerted by clinal environmental variation, particularly when sampled individuals are distributed continuously across the landscape.

Here, we used a dataset of nearly 2 million common SNPs identified by whole-genome sequencing of 202 accessions from the highly selfing annual legume Medicago truncatula to identify and investigate the population genetics of the molecular targets of adaptation to climate. We used mixed linear models to identify SNPs associated with the three climate variables that
capture the greatest proportion of variance in climate across the range of *M. truncatula*—annual mean temperature (AMT), precipitation in the wettest month (PWM), and isothermality (ITH).

This approach is conceptually similar to the logistic regression analyses of Joost et al. (2007) or Yang et al. (2012) except that we treated the environment, rather than the SNP genotypes, as the response variable. Using the environment as the response variable and SNPs as potential explanatory variables facilitated controlling for confounding effects of population structure using a kinship (K-) matrix of relatedness as a covariate. While the climate experienced by a plant accession is certainly not caused by its genotype, we assume that climate variation is a proxy for one or more unknown phenotypes that mediate adaptation to climate. Our analysis can therefore be considered a “reverse ecology” approach to identify loci that may underlie adaptation to clinal climate variation without prior knowledge of specific traits mediating that adaptation (Ellegren 2008; Li et al. 2008).

We examined the 100 SNPs showing the strongest association with each climate variable as candidates underlying adaptation to climatic conditions. We used BLASTX (Altschul et al. 1990) searches against the *Arabidopsis thaliana* genome, to see whether candidate SNPs tagged genome regions with high sequence similarity to *A. thaliana* genes having documented roles in climate adaptation phenotypes. We also grew an independent sample of *M. truncatula* accessions in controlled conditions representing extremes of temperature and soil moisture in the native range, and tested for a correlation between genotypes at candidate SNPs and performance in the growth experiment. Finally, using nucleotide diversity in genomic windows surrounding candidate SNPs to characterize the selective history of candidate regions we tested the hypothesis that candidate SNPs have evolutionary histories differing from the rest of the genome. We found that candidates for AMT and PWM are significantly more likely to lie in genes than
randomly drawn reference SNPs. In comparison to regions flanking a reference sample of 10,000 randomly-drawn SNPs, the genomic regions flanking candidate SNPs for AMT and PWM harbor less diversity, have increased differentiation between haplotypes carrying alternate alleles, and harbor an excess of common haplotypes. These patterns are consistent with soft, partial sweeps and temporally unstable selection acting on polygenic traits.

**METHODS**

**Genomic data collection:** Genomic SNP data were collected by the *Medicago truncatula* HapMap Project (MHP; www.medicagohapmap.org) and sequencing methods are described in full by Branca et al. (2011) and Stanton-Geddes et al. (2013). Briefly, we used Illumina to sequence genomes of 202 *M. truncatula* accessions collected from across the native range (Figure 1). We aligned the sequence data to the *M. truncatula* reference genome (Mt 3.5 release, Young et al. 2011) to identify SNPs within the sample. For the analyses presented here, we filtered the SNP data to include only biallelic sites for which there was sequence coverage in at least 100 accessions, and for which the less-common allele was present in at least 10% of sampled accessions (minor allele frequency, MAF, $\geq 0.10$), leaving 1,918,637 SNPs. We also used the Mt 3.5 reference genome annotation to determine whether SNPs were in coding regions or intergenic space. All genetic data are available from www.medicagohapmap.org, and Illumina reads are deposited in the NCBI Short Read Archive (SRP001874).

**Climate data acquisition and processing:** We conducted spatial analysis of climate variables in the statistical computing environment R (v 3.0.2; R Core Team 2013) using packages for species distribution modeling and geospatial analysis (maps, dismo, and their dependencies).
Within R, we downloaded raster-formatted climate data from the Worldclim database (www.worldclim.org; Hijmans et al. 2005), and extracted values for all 19 of the BioClim summary variables at the collection sites of the 202 MHP accessions included in this study plus 24 accessions identified as belonging to a diverged subgroup (Branca et al. 2011; Stanton-Geddes et al. 2013), and 671 M. truncatula collections recorded in the Global Biodiversity Information Facility (GBIF; www.gbif.org). We included this larger sample of collection localities to obtain a more complete description of the climate conditions experienced by the broader population of M. truncatula from which the MHP accessions were sampled. Regardless, the specific variables identified as describing the most variation in climate were very similar whether we used the 671 location sample or information for only the 202 re-sequenced accessions included in the genomic analyses.

To identify the subset of climate variables that best summarize the range of environments occupied by M. truncatula, we performed principal components analysis on the 19 BioClim variables at all MHP and GBIF locations. In order to facilitate the biological interpretation of the climatic gradients we then selected for subsequent analyses the single BioClim variable that most strongly loaded on each of the first four principal component axes, which together captured 91.4% of variation in the climate data (Table S1). These climate variables were annual mean temperature (AMT), precipitation in the wettest month of the year (PWM), precipitation in the coldest quarter of the year (PCQ), and isothermality (ITH; the mean diurnal range of temperatures as a percentage of the annual range of temperatures). Because the two precipitation variables, PWM and PCQ, are highly correlated (Figure S1; correlation = 0.94, P < 0.001), we report most results for only AMT, PWM, and ITH.
Identification of climate-associated SNPs: The 202 accessions in our sample are each assigned to a different collection location, so that they are effectively distributed continuously across the landscape. For such a sampling scheme, analytical frameworks that rely on characterizing within and among-population diversity or differentiation may be less effective at identifying targets of selection (De Mita et al. 2013). Therefore, we identified candidates of adaptation to climate by fitting, for each SNP, a mixed linear model predicting climate values with SNP genotype controlled for possible confounding effects of population structure. We tested for associations between the genotypes at all 1,918,637 SNPs, and the value of each of the three climate variables at the collection sites for the 202 accessions, using the mixed linear model method implemented in Tassel (v 3.0; Bradbury et al. 2007; Zhang et al. 2010). Inspection of quantile-quantile (Q-Q) plots for the distribution of association p-values from a first round of the mixed linear model analysis revealed a strong excesses of highly-associated markers in AMT and PWM; we therefore transformed the raw Bioclim values for these variables using a Box-Cox procedure, using longitude as a covariate, resulting in a distribution of P values similar to expectations (based on Q-Q plots).

To reduce the chance of identifying false associations due to population structure, we conducted association analysis with only those SNPs that had MAF $\geq$ 10%, and used a kinship (K-) matrix as a covariate (Yu et al. 2006; Bradbury et al. 2007; Zhang et al. 2010). We estimated the K-matrix from 40,000 randomly sampled SNPs (5,000 from each of $M.\ truncatula$’s eight chromosomes). When we ran the analyses using simple linear models without the K-matrix covariate, we found a strong excess of highly-associated markers, as defined by P values, for all three climate variables, an expected effect of confounding created by the spatial auto-correlation of population structure and variation in climate (Meirmans 2012; De Mita et al. 2013; Guillot
and Rousset 2013); but this effect disappeared when we included the K-matrix covariate in the analysis (Figure S2). Many candidate SNPs identified using the K-matrix-controlled mixed linear model, however, were also among the top-associated SNPs identified by simple linear models (data not shown).

For each of the three climate variables, we identified the 100 SNPs with the highest \(-\log_{10}(p)\) values for association, and for subsequent analyses considered these as candidates underlying adaptation to that climate variable (Table S2). The focus on the 100 SNPs with smallest P-values—rather than those meeting a pre-determined P-value cutoff—may increase the risk of including false positives in the pool of candidates while at the same time likely missing many genes, perhaps thousands that contribute to variation in polygenic traits. If we had focused only on SNPs having effects below a stringent P-value cutoff we would have risked excluding loci of smaller effect. The net effect of false positives among the top 100 candidates should be to weaken evidence for selection acting on the candidate SNPs as a group, because false positives should have population-genetic characteristics that are similar to the genomic background. Thus, our more inclusive candidate pool makes our subsequent tests for differences between candidate SNPs and randomly-selected reference loci more conservative.

To better understand the explanatory power of the top 100 SNPs, we used SNP genotypes at each of the candidate loci as predictor variables in multiple linear models fitted to the climate variables, then ran a model comparison analysis (the stepwise AIC procedure implemented as the R function “stepAIC”) to determine how many candidate SNPs were retained in the best-fit model. For all three climate variables, the best-fit multiple regression model explained approximately 90% of the variation (adjusted $R^2 = 0.93$ for AMT, 0.86 for PWM, and 0.91 for
ITH), and retained more than 90% of the top 100 SNPs (96 for AMT, 93 for PWM, and 98 for ITH).

To search for potential functional roles of genes containing candidate SNPs, we used BLASTX to search the TAIR database (www.arabidopsis.org) for *A. thaliana* homologs of genes containing candidate SNPs. We then examined the functional annotation for the identified *Arapbidopsis* homologs. Functional annotation for *Arabidopsis* is far better developed than that for *M. truncatula*, and much of the work on the functional genetics of *Medicago* has focused on genes related to the symbiosis with rhizobia, so the *Arabidopsis* database provided a richer resource for understanding possible functions of our candidates.

**Experimental validation:** To see whether candidate SNPs contribute to performance in specific climate conditions, we tested the performance of an independent sample of *M. truncatula* accessions—68 lines in the MHP panel that were excluded from the climate association because their original collection sites are unknown (Table S3)—in conditions intended to replicate the upper range of temperature and rainfall experienced by natural populations. We scarified seeds by agitation with sterile sand, cold-stratified them for 4 days at 4°C, germinated them on sterile filter paper, and planted germinated seeds in Sunshine Mix LP5, in 1.5-inch cone-tainers. When plants grew their second adult leaf, we introduced them into a climate-controlled growth chamber set to a 14:10 light-dark cycle and 35°C—towards the upper end of the distribution of summer temperatures in the native range. Plants in the growth chamber were bottom-watered, maintaining a higher level of soil moisture than experienced by most natural populations. We harvested all plants on the same day, after they had spent an average of 25.5 days (full range: 9 to 29 days) in the growth chamber conditions, and measured their height
as the length of the longest branch, following Stanton-Geddes et al (2013). We tested for a relationship between the number of candidate alleles associated with high AMT and high PWM in each of the 68 experimental lines and their growth rate in the growth chamber (as height gained per days in the chamber).

**Evaluating signs of selection near climate-associated SNPs:** To understand how selection has shaped diversity in genomic regions flanking candidate SNPs we examined patterns of nucleotide diversity in 10,000-base (10kb) regions centered on each of the 100 candidate SNPs for each climate variable. For each 10kb region we calculated two measures of nucleotide diversity: the number of segregating sites per site, $\theta_W$ (Watterson 1975), average pairwise differences per site, $\theta_\pi$ (Nei 1987) in the entire sample (“total”), the haplotypes carrying the less common (minor) allele, and the haplotypes carrying the more common (major) allele at the candidate SNP. We characterized differentiation between the 10kb windows flanking the major and minor alleles at each candidate SNP using $F_{ST}$ (Wright 1940), which we calculated as the mean of $F_{ST}$ values calculated for each SNP segregating in that window.

Finally, we tested the hypothesis that regions containing the candidate SNPs had recently experienced a selective sweep using the H12 statistic of Garud et al. (2013). H12 is a haplotype-based statistic that is equal to the square of the sum of the frequencies of the most-common and second-most-common haplotypes, plus the sum of the squares of the frequencies of each other unique haplotype. The statistic scales between 0 and 1, with values closer to 1 indicating that one or two haplotypes dominate the sample, consistent with the effects of a recent selective sweep. The H12 statistic has the advantage that it can robustly detect both classical “hard” sweeps, in which a single haplotype dominates the sample, as well as “soft” sweeps of multiple favored
haplotypes, as may occur if multiple haplotypes carry alleles that are advantageous in the same way, or if alternate haplotypes are adaptive in different parts of the region sampled (Hermisson and Pennings 2005; Garud et al. 2013). To calculate $H_{12}$, we reconstructed haplotypes for the 10kb window flanking each candidate SNP using fastPhase (version 1.2; Scheet and Stephens 2006), then calculated $H_{12}$ from the fastPhase-estimated haplotype frequency spectrum.

We compared the median and mean of the summary statistics calculated ($\theta_W$ total, $\theta_W$ minor, $\theta_W$ major, $F_{ST}$, and $H_{12}$) for the 10kb regions flanking candidate SNPs to summary statistics calculated on regions flanking a genome-wide reference sample of 10,000 randomly drawn reference SNPs. For two climate variables (AMT and PWM) the candidate SNPs were enriched for genic regions (Table 1; exons and introns, but not genic regions annotated as being parts of transposable elements in the Mt3.5 genome assembly; Young et al. 2011). If the summary statistics we use to describe nucleotide variation differ between genic and intergenic regions—which is seen in the Medicago genome (Branca et al. 2011; Table 1, below), presumably due to differential effects of purifying and background selection, then randomly sampled reference SNPs are not an appropriate standard for comparison. For this reason, we compared the summary statistics for candidates for each of the three climate variables to both the randomly sampled reference SNPs, and to reference SNPs conditioned on whether they are annotated as genic (coding sequence and introns) or intergenic.

All analyses, except for the fastPhase haplotype reconstructions used to calculate $H_{12}$, were conducted in R. Once we had calculated the diversity and differentiation statistics for regions flanking climate candidate SNPs and the reference sample of 10,000 SNPs, we compared the statistics observed for the candidate SNPs to expectations estimated from the background sample using Wilcoxon sign-rank tests for differences of medians.
RESULTS

Climate data and candidate selection: The 202 accessions of *Medicago truncatula* we analyzed were originally collected at sites from western Morocco to central Turkey (Figure 1; approximately 5,000 kilometers east to west) and from southern France to the African Mediterranean Coast (approximately 2,000 kilometers north-south). Climatic conditions across this range separated along four principal components (PCs) that together accounted for 91.4% of variation in the climactic conditions from which our accessions were originally collected. In order for our results to have clearer biological interpretation, we worked with the Bioclim variables having the largest loadings on each of the first four PCs rather than the PC axes themselves.

Annual mean temperature (AMT) and isothermality (ITH) were the major loadings on PC1 and PC4, accounting for 45.6% and 7.5% of total variation, respectively. The major loadings on PC2 and PC3 (25.0% and 23.5% of total variation, respectively) were precipitation in the wettest month (PWM) and precipitation in the coldest quarter (PCQ). Although PCs are orthogonal to one another, there are significant correlations among the climatic variables that are the most important determinant of each PC (Figure S1); AMT and ITH are significantly correlated ($\rho = 0.45$, $P < 0.001$) and PWM and PCQ are significantly correlated ($\rho = 0.95$, $P < 0.001$). Because of the very high correlation between PWM and PCQ, we conducted subsequent analyses on PWM only.

For each of the three climate variables, we identified the 100 SNPs with strongest association to climate as candidates potentially underlying adaptation to each climate variable (Table S2). All candidate SNPs had $-\log_{10}(p)$ values $> 4.2$, with median $-\log_{10}(p)$ values of
4.68 for AMT candidates, 4.88 for PWM, and 4.47 for ITH. The minor allele frequency (MAF) of candidate SNPs for AMT and ITH did not differ significantly from the MAF of randomly sampled reference SNPs (P = 0.23 and 0.12, respectively, in Wilcoxon sign-rank tests); but the median MAF of candidates for PWM is significantly lower than the median MAF for reference SNPs (P = 0.0093). We observed that candidate SNPs for AMT and PWM were more likely to lie in annotated genes than expected by chance: 51% of AMT candidates and 41% of PWM candidates, compared to 31.7% of the reference sample (P < 0.0001 and P = 0.031, respectively, based on 10,000 random samples of 100 SNPs from the reference set of 10,000 SNPs). By contrast, just 30% of ITH candidates are genic, which is not significantly different from the reference sample (P = 0.68). Of AMT candidates, 8% are in annotated transposable elements, which is not significantly different from the reference sample (P = 0.42). None of the PWM candidates are in transposable elements, which is not seen in any randomly drawn set of 100 reference SNPs. Of the ITH candidates, 13% are in transposable elements, significantly more than expected based on the reference sample (P = 0.024).

**Functions of candidate genes:** Among the 300 candidate SNPs, 122 were within annotated genic regions (coding sequence or intron) and these 122 SNPs tagged 62 unique genes in the Mt3.5 genome annotation (Table S2; Young et al. 2011). BLASTX searches of the A. thaliana genome identified homologs for 42 of the M. truncatula genes containing candidate SNPs. Several of the A. thaliana homologs have been subjects of empirical study that established functional roles that may be important in adaptation to the abiotic environment or in defense against herbivores of pathogens (Table 2).
Genomic distribution and linkage of candidate SNPs: Candidate SNPs for all three climate variables were distributed genome-wide, with sites on every chromosome tagged (Figures S3-5). However, for each climate variable there were clusters of candidate SNPs seen as “peaks” in Manhattan plots (Figures S3-5). Pairwise plots of linkage disequilibrium (LD) among candidates (Figures S6-8) reveal that candidate SNPs in these clusters are more strongly linked than candidate SNPs in general. We also expected that candidates for the same climate variable would tend to be in stronger LD with one another than non-candidates, simply because the alternate alleles of candidate SNPs are, by definition, all strongly associated with the same candidate variable. Indeed, the median D’ estimate for each of the three sets of candidates was greater than the median of the median D’ for 30 randomly-drawn sets of 100 reference SNPs; and the median linkage among PWM and ITH candidates was greater than the median for any of the randomly-drawn sets of reference SNPs (Figure 2).

Experimental validation: We found positive relationships between the number of alleles associated with high AMT and PWM carried by each of the 68 lines used in the growth chamber experiment, and their rate of growth in the high temperature and high soil moisture conditions (Figure S9). The relationship between alleles associated with high AMT at AMT candidate loci and growth rate was positive but not greater than expected by chance (ρ = 0.13, P = 0.29); and there was a positive, statistically significant relationship between growth rate and candidate alleles associated with high PWM (ρ = 0.40, P = 0.0008). A linear regression between the number of “high” alleles at AMT and PWM candidates together revealed an even stronger positive relationship (ρ = 0.42; regression adjusted R² = 0.16, P = 0.0004; Figure S9).
Diversity and differentiation in regions flanking candidates: We compared summary statistics from 10kb segments of sequence centered on each candidate SNP to statistics from 10kb segments centered on each of the 10,000 randomly drawn reference SNPs using one-tailed Wilcoxon sign-rank tests. Because candidate SNPs for ATM and PWM differed from the reference sample in the proportion of genic SNPs, we also separately compared genic candidates to genic reference SNPs, and intergenic candidates to intergenic SNPs. We found that comparisons based on $\theta_W$ and $\theta_\pi$ were qualitatively similar ($\theta_\pi$ reported in Table S4).

Regions flanking candidates for all three climate variables differed significantly from regions flanking reference SNPs, though not in the same manner. Compared to the reference sample, candidate SNPs for AMT had reduced nucleotide diversity in haplotypes carrying the minor allele ($\theta_W$, minor; Table 1, $P = 0.0041$), and greater differentiation between backgrounds carrying the alternate alleles ($F_{ST}$; Table 1, $P = 0.0064$). They had also experienced stronger, or more recent, selective sweeps, as indicated by elevated $H_{12}$ (Table 1, $P = 0.032$). We found similar and stronger patterns for PWM candidates. They had reduced nucleotide diversity in all backgrounds considered together ($\theta_W$, total; Table 1, $P < 10^{-4}$), in major allele backgrounds ($P < 10^{-4}$), and in minor allele backgrounds ($P < 10^{-4}$), as well as elevated differentiation between backgrounds ($P = 0.045$) and greater haplotype frequency distortion ($P < 10^{-4}$). The differences between AMT and PWM candidates compared to reference SNPs were largely due to differences in regions flanking genic candidates. Intergenic AMT candidates were not significantly different than intergenic reference regions for any of the summary statistics (all $P > 0.25$, Table 1). Intergenic PWM candidates did not differ from reference regions in either $F_{ST}$ or $H_{12}$ (both $P > 0.2$, and differences in diversity between intergenic PWM candidates and intergenic reference regions were both smaller and statistically less significant (Table 1, Figure 3).
In contrast to AMT and PWM candidates, regions flanking ITH candidates had greater nucleotide diversity, similar differentiation between minor and major allelic backgrounds ($F_{ST}$), and lower than expected frequencies of common alleles (i.e. lower H12, Table 1, Figure 3), compared to regions flanking the 10,000 randomly drawn reference SNPs. Moreover, unlike AMT and PWM candidates (where differences from reference sample were driven by genic SNPs) differences between ITH candidate and reference sample were primarily due to SNPs in annotated intergenic regions.

**DISCUSSION**

Local and clinal adaptation, which is pervasive in natural populations, has been a long-term focus in evolutionary ecology, and is probably important in shaping intraspecific diversity. We used mixed linear models to search nearly 2 million single-nucleotide polymorphisms assayed in more than 200 accessions of *Medicago truncatula* for candidate loci that may be responsible for local adaptation along clines in temperature, precipitation, and seasonality (Bradbury et al. 2007; Zhang et al. 2010). This approach let us scan for genome regions that might underlie adaptation to climate without making assumptions as to which phenotypes mediate this adaptation. The candidates we identified tagged annotated genes with high sequence similarity to genes with roles in climate adaptation and pathogen defense in *Arabidopsis thaliana* and other taxa. For two of the three climatic gradients, AMT and PWM, candidate SNPs were significantly enriched for genic regions, and candidates for these variables predicted performance in an independent sample of accessions grown in controlled conditions (Figure S9). Finally, candidates for all three lay in regions with patterns of nucleotide diversity that differ from genomic expectations (Table 1; Figure 3).
Identification of candidates for adaptation to climate: The mixed linear model approach we used to identify candidate genes underlying adaptation along environmental clines is computationally efficient for large datasets, and flexible enough to allow the incorporation of biologically meaningful covariates to control for population structure. This approach can, however, risk identifying false positives due to the spatial correlation of population structure with climate (Coop et al. 2010; Meirmans 2012; De Mita et al. 2013; Günther and Coop 2013). This problem is similar to the difficulty of identifying predictors of morphological variation or species assemblage composition from collections of spatially correlated environmental variables, which has been a major concern of landscape ecologists for decades (Smouse et al. 1986; Legendre 1993; Rousset 1997). As in these other tests for causal factors among spatially correlated data, searches for clinal adaptation based on $F_{ST}$ outliers or strong covariance between the environment and within-population SNP frequency can be troubled by false positives (Narum and Hess 2011; Keller et al. 2012) and may be less sensitive to the effects of weak selection (De Mita et al. 2013).

To minimize the probability of identifying SNPs that show associations with the environment due to demographic history rather than selection, we included a K-matrix as a covariate in our analyses (Bradbury et al. 2007; Zhang et al. 2010) and we limited our candidate pools to SNPs with minor allele frequencies > 0.10. False positives among candidates also are possible because we did not rely on a stringent P-value cutoff but rather considered the 100 SNPs with the strongest association with climate variables as candidates. At the same time, if adaptation to climate is mediated by polygenic traits (Messer and Petrov 2013), we expect individual loci to have small effects, and a more stringent P-value cutoff may eliminate many
real candidates from the candidate pool. Indeed, even though we expect false positives, candidate SNPs for AMT and PWM are significantly enriched for genic regions and thus are clearly not a random sample of genomic variation. Under the assumption that the majority of selectively important variation lies in genic regions, enrichment of candidate SNPs in protein coding genes is consistent with candidates representing true targets of adaptation to climate (Hancock et al. 2010, 2011). This enrichment of genic regions does not appear to be an artifact of applying the mixed linear model to genomic data, since the annotation of ITH candidate SNPs does not differ from that of random samples.

**Functional roles of candidates:** Candidate SNPs fell within 62 unique genes, according to the annotation of the Mt3.5 assembly of the *M. truncatula* genome (Table S2; Young et al. 2011). Many of these genes were putative orthologs of *A. thaliana* genes with established function in phenotypes important for adaptation to climatic variation. Of the 62 genes directly tagged by candidate SNPs there are four pairs of apparent paralogs that show greatest sequence similarity to the same *A. thaliana* gene (Table 2). Three of these pairs encode proteins for which functional information from *A. thaliana* strongly support a direct role in plant defense, including the *EFR* gene coding for a receptor for the bacterial PAMP EF-Tu, a TIR-NBS-LRR class disease resistance protein, and the F-box encoding protein *CPR*. The *EFR* gene was identified by Fournier-Level et al. (2011) as a potential contributor to climate-specific survival in *A. thaliana*, and an *M. truncatula* homolog of the *A. thaliana CPR30* has been shown to confer resistance to the oomycete *Aphanomyces euteiches* (Bonhomme et al., 2014). Bonhomme et al. also speculated that this *M. truncatula* F-box gene may play a direct role in response to abiotic stress. Defense-related genes may be involved in adaptation due to indirect effects of temperature and
precipitation on pathogen occurrence and pathogen pressure can be strongly affected by
temperature and precipitation. The homologs of several other *Medicago* genes tagged by
candidate SNPs also have established functions that are consistent with direct roles in mediating
response to temperature, drought, or pathogens—all of which are likely to be affected either
directly (thermotolerance and water availability) or indirectly (pathogen pressure) by the climate
variables we studied (Table 2).

We also found that alleles associated with higher temperature and rainfall at AMT and
PWM candidates predicted the performance of an independent sample of plant lines grown in
controlled conditions representing the upper range of temperature and soil moisture found in the
native range (Figure S9). This is consistent with an adaptive role for these alleles, or variants
linked to them, in relation to temperature and precipitation. Evaluation of the candidates could be
further bolstered if we knew the specific phenotypes that are responsible for adaptation to the
climatic variables we studied, assuming that populations are locally adapted to climatic
conditions.

Although empirical data on traits that mediate adaptation to climate at the range-wide
scale are not available, flowering time is a strong *a priori* trait of interest, based on both the
importance of phenology to local adaptation in plant species (Bradshaw and Holzapfel 2001;
Mimura and Aitken 2010; Keller et al. 2012; Colautti and Barrett 2013) and evidence for
adaptive divergence in flowering time between two *M. truncatula* populations from southern
France (Bonnin et al. 1996). Based on previous greenhouse measurements of the accessions in
our sample (Stanton-Geddes et al. 2013), flowering time is correlated with the climatic variables
we examined, though relatively weakly (with AMT, $\rho = -0.36$, linear regression $R^2 = 0.13$, $P <$
0.001; and $\rho = -0.22$, $R^2 = 0.05$, $P = 0.004$ with ITH; Figure S1). Given the magnitude of these
correlations, it may not be surprising that there is no overlap between flowering time candidate loci identified by GWA (Stanton-Geddes et al. 2013) and the candidate loci we identified. Another phenotype likely mediating adaptation to climate in *M. truncatula* is drought tolerance, which is understood to be important for clinal adaptation in Mediterranean habitats (Grivet et al. 2011) and we did identify several climate candidates with possible roles in drought tolerance (Table 2). However, no functional genetics or GWAS study has independently identified loci underlying drought tolerance in *M. truncatula*, so we have no independent confirmation for a role of drought tolerance.

**Genetic architecture of adaptation to climate:** Although the candidate loci we identified are spread across the genome, many candidate SNPs are found in spatial clusters within which candidates are in high linkage disequilibrium (LD, Figures S3-8). A particularly striking case is the roughly 2.5 Mb region on chromosome 2 containing a large concentration of tightly linked PWM candidates, a single ITH candidate, and several homologs of the *A. thaliana* CPR gene (Figure 4; Figures S4 and S7). The strong LD among candidates within this region suggests that this region may contain a large inversion. Chromosomal inversions can maintain sets of co-adapted alleles at multiple loci, and have been implicated in adaptation to climate in other plants, including *Mimulus* (Lowry and Willis 2010), *A. thaliana* (Long et al. 2013) and teosinte (Fang et al. 2012; Pyhäjärvi and Hufford 2013).

Spatial clustering of candidate SNPs in high LD with one another (Figure 2) is consistent with selection operating on these regions. Spatial clustering is expected if multiple adaptive substitutions occur in the same gene region, and elevated LD between physically unlinked markers can be created when multiple loci contribute to variation in a phenotype under selection.
(Berg and Coop 2013). However, these outcomes also mean that our candidate sets are nonrandom samples of 100 SNPs, which may contribute to differences in diversity and differentiation for genome segments flanking candidate SNPs, compared to randomly sampled reference SNPs. Indeed, when we filtered our sets of 100 SNPs to include only candidates at least 20kbp apart, we found that these reduced sets of “independent” candidates differed less significantly from the reference SNPs—although this may also be an effect of smaller sample size.

**Nucleotide diversity and differentiation:** Candidate SNPs for adaptation to AMT and PWM, but not ITH, are enriched for genic regions (Table 1). This is expected if adaptation primarily proceeds through changes in protein-coding regions (Hancock et al. 2011). The AMT and PWM candidates also lie in regions harboring less nucleotide diversity, greater differentiation between alternate alleles, and an enrichment of common haplotypes (elevated H12), compared to randomly drawn reference SNPs (Table 1). Candidates for ITH also lie in regions differing significantly from the reference sample, but in the opposite direction: ITH candidates are in regions with greater nucleotide diversity, have lower H12 than the reference, and show no elevated differentiation between allelic backgrounds (Table 1).

Because *M. truncatula* genes harbor lower diversity and presumably experience stronger purifying selection than non-coding regions (Branca et al. 2011), the candidate SNPs for AMT and PWM may differ from randomly selected reference SNPs simply because of the greater representation of genic SNPs among the candidates. However, separating the candidate to reference comparison on the basis of whether SNPs were in annotated genes revealed that differences between both AMT and PWM candidates compared to reference SNPs are due almost
entirely to regions flanking genic SNPs. Genic AMT candidates show reduced diversity and increased differentiation relative to genic reference SNPs whereas intergenic candidates did not differ from intergenic reference SNPs (Table 1, Figure 3). Similarly, genic PWM candidates differed from genic reference SNPs more strongly than intergenic candidates differ from intergenic reference SNPs (Table 1, Figure 3). In contrast, differences between ITH candidates and the reference sample are due to intergenic regions (Table 1, Figure 3).

The lower nucleotide diversity, especially in the minor allele background, seen for AMT and PWM candidates may be due at least in part to stronger than average purifying selection, which is expected when two separate allelic classes are favored under different climatic conditions and mutations in either allelic background are selectively disadvantageous (Charlesworth 1998). Alternatively, lower diversity in the minor allele background may reflect partial sweeps due to recent adaptation. This seems possible for both the PWM and AMT candidates as the minor allele harbors significantly less diversity than expected (θW, minor, in Table 1) and an excess of common haplotypes as reflected in elevated H12. Both locally restricted purifying selection and local sweeps would be consistent with the elevated differentiation we see between the alternate alleles of genic candidates for AMT and PWM (FST, Table 1). Lower nucleotide diversity in regions flanking SNPs most strongly associated with climate variation is, however, inconsistent with adaptation that maintains two alternative alleles through balancing selection (Nordborg et al. 1996; Hedrick 2006). In contrast to patterns seen in candidates for AMT and PWM, regions containing the candidates for ITH have greater nucleotide diversity than expected. Elevated diversity may be consistent with local or clinal adaptation maintaining higher diversity (i.e. balancing selection) however, the minor and major
allele backgrounds show no evidence of elevated differentiation between allelic backgrounds, which is also expected under that scenario (Table 1; Nordborg et al. 1996).

CONCLUSIONS

Overall, the patterns of diversity we find among the candidate loci are consistent with soft or partial sweeps, in which adaptation is driven by small shifts in allele frequency at multiple loci. This may reflect an unstable selective environment (Coop et al. 2009; Pritchard et al. 2010) due to changing climates or recent expansion of the species’ geographic range, or it may reflect ongoing adaptation fueled by standing variation and recurrent mutation (Hermisson and Pennings 2005; Messer and Petrov 2013). The homology of our candidates to functionally important A. thaliana genes, and the results from the growth-chamber experiment, which showed significant positive relationship between the number of PWM and AMT candidates and growth in a stressful environment, provide support that the candidates we identified contribute to adaptation to climatic environment. Transplant experiments to directly test for local adaptation along the climatic gradients we examined could confirm the selective significance of the range of climate variations we examine and could also identify specific phenotypes that are associated with adaptation to climate.

Local adaptation can leave a signature of balancing selection, with strong differentiation between alleles favored in distinct habitats and elevated diversity in species-wide samples (Lewontin and Krakauer 1973; Charlesworth 1998; Hedrick 2006). However, despite the pervasiveness of local adaptation, there are few empirical examples of locally adapted genes showing strong balancing selection (Hedrick 2006; Lee and Mitchell-Olds 2012). In this study we used whole-genome data, so the lack of evidence for strong balancing selection on genes
responsible for adaptation along environmental clines, a form of local adaptation is unlikely to be
the result of failing to examine phenotypically important loci. It is possible that we fail to find
evidence for balancing selection because balancing selection is very difficult to detect for genes
underlying polygenic traits, and adaptive traits are often polygenic (Holeski and Kelly 2006;
Chevin and Hospital 2008; Messer and Petrov 2013). However if this were the reason we would
not expect to find that diversity in the genomic regions flanking candidate SNPs differ from the
rest of the genome. We therefore conclude that our results are most consistent with a history of
adaptation to climate variation via selection on polygenic traits, which proceeded in locally
restricted soft sweeps. As genomic data from range-wide samples become available for more
species, it will be possible to determine if soft sweeps are the typical mode of adaptive evolution
at genes underlying adaptation to climatic gradients.

ACKNOWLEDGEMENTS

We thank the Noble Foundation, Joelle Ronfort, Jean-Marie Prosperi, Laurent
Gentzbittel, Sergey Nuzhdin, and Katy Heath for providing seed, and Roxanne Denny, the Noble
Foundation, and INRA at Montpelier for maintaining M. truncatula germline collections. Lea
Graber collected measurements in the growth chamber experiment. Jeffrey Ross-Ibarra and one
anonymous reviewer provided valuable comments on the manuscript. Computational resources
provided by the Univ. of Minnesota Supercomputing Institute greatly facilitated data analyses.
The work was funded by National Science Foundation Grants 0820005 and 1237993.
REFERENCES


Yoder et al: Genomics of adaptation to climate


1 Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: molecular population genetics of

3 Hijnman, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution

6 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.
7 PLoS Genet. 6:e1000862.


12 approaches To detect loci under selection in clinal populations. Evolution (N. Y). n/a–n/a.

14 spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape

16 Keller, S. R., N. Levsen, M. S. Olson, and P. L. Tiffin. 2012. Local adaptation in the flowering-

18 Kelly, J. K., B. Koseva, and J. P. Mojica. 2013. The genomic signal of partial sweeps in Mimulus

20 Kujala, S. T., and O. Savolainen. 2012. Sequence variation patterns along a latitudinal cline in


26 3:e4010.

27 Levy, A., M. Erlander, M. Rosenthal, and B. L. Epel. 2007a. A plasmodesmata-associated beta-

29 Levy, A., D. Guenoune-Gelbart, and B. L. Epel. 2007b. β-1,3-glucanases: Plasmodesmal gate


Table 1. Summary statistics for candidate SNPs, reference SNPs, and 10kb windows centered on each.

<table>
<thead>
<tr>
<th>Marker set</th>
<th>N</th>
<th>θ&lt;sub&gt;W&lt;/sub&gt;, total</th>
<th>θ&lt;sub&gt;W&lt;/sub&gt;, major</th>
<th>θ&lt;sub&gt;W&lt;/sub&gt;, minor</th>
<th>F&lt;sub&gt;ST&lt;/sub&gt;</th>
<th>H12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Candidates for</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMT: Top 100</td>
<td>100</td>
<td>0.0017 (0.610)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.0018 (0.579)</td>
<td>0.0011 (0.0041)</td>
<td>0.3799 (0.006)</td>
<td>0.0058 (0.032)</td>
</tr>
<tr>
<td>Genic</td>
<td>51</td>
<td>0.0016 (0.052)</td>
<td>0.0016 (0.402)</td>
<td>0.0010 (0.0006)</td>
<td>0.4312 (&lt; 0.001)</td>
<td>0.0126 (0.003)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>41</td>
<td>0.0019 (0.550)</td>
<td>0.0019 (0.653)</td>
<td>0.0011 (0.2683)</td>
<td>0.3414 (0.692)</td>
<td>0.0039 (0.529)</td>
</tr>
<tr>
<td><strong>PWM: Top 100</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genic</td>
<td>41</td>
<td>0.0012 (&lt; 0.001)</td>
<td>0.0013 (&lt; 0.001)</td>
<td>0.0004 (&lt; 0.001)</td>
<td>0.4207 (0.035)</td>
<td>0.0201 (0.002)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>59</td>
<td>0.0013 (0.003)</td>
<td>0.0013 (0.007)</td>
<td>0.0011 (&lt; 0.001)</td>
<td>0.3401 (0.273)</td>
<td>0.0063 (0.225)</td>
</tr>
<tr>
<td><strong>ITH: Top 100</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genic</td>
<td>30</td>
<td>0.0019 (0.531)</td>
<td>0.0019 (0.526)</td>
<td>0.0017 (0.551)</td>
<td>0.3633 (0.628)</td>
<td>0.0038 (&lt; 0.001)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>57</td>
<td>0.0022 (&lt; 0.001)</td>
<td>0.0021 (&lt; 0.001)</td>
<td>0.0020 (&lt; 0.001)</td>
<td>0.3205 (0.232)</td>
<td>0.0022 (&lt; 0.001)</td>
</tr>
<tr>
<td><strong>Reference sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genic</td>
<td>3,171</td>
<td>0.0019</td>
<td>0.0019</td>
<td>0.0016</td>
<td>0.3285</td>
<td>0.0047</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0005, 0.0043</td>
<td>0.0005, 0.0045</td>
<td>0.0002, 0.0045</td>
<td>0.0851, 0.6613</td>
<td>0.0001, 0.0911</td>
</tr>
<tr>
<td>Intergenic</td>
<td>6,120</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0015</td>
<td>0.3293</td>
<td>0.0048</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0005, 0.0042</td>
<td>0.0005, 0.0044</td>
<td>0.0002, 0.0044</td>
<td>0.0851, 0.6433</td>
<td>0.0001, 0.0781</td>
</tr>
</tbody>
</table>

<sup>a</sup> Median (p-value for Wilcoxon sign-rank test of the hypothesis that the SNP set differs from the reference sample)

<sup>b</sup> First line, median value; second line, minimum and maximum for 95% of observations.
Table 2. *Medicago truncatula* gene models containing climate candidates, their *Arabidopsis thaliana* (*At*) homologs, and hypothesized function based on prior studies.

<table>
<thead>
<tr>
<th>Climate variable</th>
<th>Gene model&lt;sup&gt;a&lt;/sup&gt;</th>
<th><em>At</em> BLAST match(es)</th>
<th>Functional annotation for <em>At</em> match(es)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWM</td>
<td>1g045760.1 3-isopropylaminate dehydrogenase</td>
<td>AT4G13430 (MAM-IL)</td>
<td>methylthioalkylmalate isomerase involved in glucosinolate biosynthesis (Sawada et al. 2009).</td>
</tr>
<tr>
<td>PWM</td>
<td>2g025700.1 Calcineurin B-like protein</td>
<td>AT3G51970</td>
<td>Calcineurin B-like (CBL) proteins are calcium sensors involved in stress response signaling (Tang et al. 2012).</td>
</tr>
<tr>
<td>AMT</td>
<td>2g039300.1 protein kinase</td>
<td>AT1G73660</td>
<td>At gene encodes a putative MAPKKK and negatively regulates salt tolerance in Arabidopsis (Gao et al. 2008).</td>
</tr>
<tr>
<td>PWM</td>
<td>3g093340.1 ABC transporter ATP-binding protein/permease</td>
<td>AT1G28010 (and others)</td>
<td>Controls stomatal response to CO&lt;sub&gt;2&lt;/sub&gt; (Lee et al. 2008)</td>
</tr>
<tr>
<td>AMT</td>
<td>7g074620.1 Ethanolamine-phosphate cytycyltransferase</td>
<td>AT2G38670</td>
<td>Leaf respiration capacity during prolonged growth under short-day conditions. (Otsuru et al 2013)</td>
</tr>
<tr>
<td>ITH</td>
<td>7g092550.1 Unknown</td>
<td>AT2G01050</td>
<td>Zinc ion binding; nucleic acid binding (arabidopsis.org)</td>
</tr>
<tr>
<td>AMT</td>
<td>1g017870.1 70 kDa peptidyl-prolyl isomerase</td>
<td>AT3G25230</td>
<td>Possible role in thermal tolerance through interactions with HSP90.1 (Meiri and Breiman, 2009).</td>
</tr>
<tr>
<td>AMT</td>
<td>4g064550.1 E3 ubiquitin-ligase</td>
<td>AT3G55530</td>
<td>Drought tolerance (Zhang et al. 2008; Gao et al. 2011).</td>
</tr>
<tr>
<td>AMT</td>
<td>4g119450.1 cullin-like protein1</td>
<td>AT4G02570</td>
<td>stomatal closure under drought (Zhang et al. 2008).</td>
</tr>
<tr>
<td>AMT</td>
<td>7g082470.1 Ser/Thr protein kinase</td>
<td>AT1G70250</td>
<td>expression changes significantly under temperature stress. (Swindell et al. 2007)</td>
</tr>
<tr>
<td>AMT &amp; PWM &amp; ITH</td>
<td>2g027090.1 PDS5-like sister chromatid cohesion protein</td>
<td>AT5G47690</td>
<td>Expressed in stomata guard cells under stress (Zhao et al 2008, Obulareddy et al 2013)</td>
</tr>
<tr>
<td>AMT &amp; ITH</td>
<td>4g114850 Beta 1,3-glucanase</td>
<td>AT5G42100 (AtBG_pap)</td>
<td>Release of winter dormancy in <em>Populus</em> buds (Rinner et al 2011).</td>
</tr>
<tr>
<td>AMT</td>
<td>5g019700.1 Ser/Thr kinase</td>
<td>AT5G66710</td>
<td>Protein phosphorylation (arabidopsis.org)</td>
</tr>
<tr>
<td>AMT</td>
<td>8g077980.1 kinase ATN1</td>
<td>AT5G36930&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Disease resistance protein (TIR-NBS-LRR class) family – same AT gene is best match for both candidates</td>
</tr>
<tr>
<td>PWM</td>
<td>4g023400.1 TMV resistance N</td>
<td>AT5G20480 (EFR)</td>
<td>EF-Tu receptor PRR for PAMP elf18, immune recognition (Zipfel et al. 2006, Roux et al. 2011), and linked to climate adaptation (Fournier-Level et al 2011).</td>
</tr>
<tr>
<td>ITH</td>
<td>5g019070.1 LRR receptor-like serine/threonin kinase FEI</td>
<td>AT5G20480 (EFR)</td>
<td>EF-Tu receptor PRR for PAMP elf18, immune recognition (Zipfel et al. 2006, Roux et al. 2011), and linked to climate adaptation (Fournier-Level et al 2011).</td>
</tr>
<tr>
<td>AMT</td>
<td>5g026000.1 kinase-like protein</td>
<td>AT4G12560 (CPR30)</td>
<td>Negative regulator of defense response (Gou et al. 2009, Cheng et al. 2011) <em>M. truncatula</em> paralog (3g011020) is a major determinant of resistance to oomycete <em>Aphanomyces euteiches</em> (Bonhomme et al. 2014).</td>
</tr>
</tbody>
</table>

<sup>a</sup> International *Medicago* Genome Annotation Group (IMGAG) gene model ID in bold, with annotated gene product

<sup>b</sup> SNPs given as chromosome:position

<sup>c</sup> 10<sup>th</sup> best match, 12 with e-100 or better
$d$ AT5G36930 is the best BLAST match out of many significant hits for 4g023400.1; and the single best match for 6g074650.1.

$e$ AT4G12560 CPR30 is among 5 highest matches for Medtr7g060940
**Figure legends.**

**Figure 1.** Collection sites for the 202 lines of *Medicago truncatula* in our sample, with the native range colored according to annual mean temperature (AMT), in degrees Celsius × 10 (scale to the right of the map).

**Figure 2.** Linkage among candidate SNPs compared to null expectations. The histogram illustrates the distribution of median $D'$ values for 30 sets of 100 randomly drawn reference SNPs; median values for the top 100 candidates for each climate variable are indicated by vertical lines.

**Figure 3.** Distribution of diversity ($\theta_W$, calculated across all allelic backgrounds), divergence ($F_{ST}$ between major and minor allelic backgrounds), and sweep strength ($H_{12}$) statistics for 10kb regions centered on each of the top 100 candidate SNPs for each climate variable (colored bars; red = AMT, blue = PWM, orange = ITH), compared to 10kb regions centered on each of the 10,000 randomly drawn reference SNPs (gray bars). To control for systematic differences between genic and intergenic regions, we compare genic candidates to genic reference SNPs (left column for each statistic) and intergenic candidates to intergenic reference SNPs (right columns); Table 1 gives the number of SNPs in each of these sets. P-values in each panel are for a Wilcoxon sign-rank test of the hypothesis that the median statistic for the candidate set is greater or less than the median for the reference set.

**Figure 4.** Manhattan plot of the tightly-linked cluster of PWM candidate SNPs on chromosome 2. Individual points plot the $-\log_{10}(p)$-value for individual SNPs. Points representing SNPs in the
top 1,000 candidates (across the whole genome) are colored according to their linkage disequilibrium (D’) with the highest-associated SNP in the illustrated region (at position 8,057,576; indicated with a red *): red = higher values of D’, shading to yellow = lower.
Linkage among sets of SNPs

Random sets of 100 SNPs

100 top SNPs for AMT

ITH

PWM