Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula

Antoine Branca1,*, Timothy D. Paape1, Peng Zhou1, Roman Briskine2, Andrew D. Farmer3, Joann Mudge3, Arvind K. Bharti3, Jimmy E. Woodward4, Gregory D. May5, Laurent Gentzbittel3, Cécile Ben3, Roxanne Denny1, Michael J. Sadowsky1, Joëlle Ronfort3, Thomas Bataillon3, Nevins D. Young1,2, and Peter Tiffin1,2

*Department of Plant Biology, University of Minnesota, Saint Paul, MN 55108; 1Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108; 2Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455; 3National Center for Genome Resources, Santa Fe, NM 87505; 4Université de Toulouse, Laboratoire Symbiose et Pathologie des Plantes, ENSET, 31326 Castanet-Tolosan, France; 5Department of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN 55108; 6Unité Mixte de Recherche AGAP, Institut National de la Recherche Agronomique Montpellier domaine de Melguell, F-34139 Mauguio, France; and 7Bioinformatics Research Center and Institute of Biology, Aarhus University, 8000 Aarhus, Denmark

AUTHOR SUMMARY

Sequencing the genomes of multiple individuals of a single species, a field known as population genomics, provides the opportunity to extend our understanding of mutation, recombination, and selection in shaping genomic diversity at the species level. Beyond evolutionary processes, the sequencing data from a population sample are important for developing the tools and resources needed for genome-wide association studies (GWAS). The use of GWAS, especially in humans, for identifying genetic variants that influence or have profound effects on complex traits, such as disease susceptibility, height, and yield, remains challenging (1). However, in the case of plant species where data on various traits can be collected in highly replicated experiments with controlled environmental conditions, GWAS seems to be a potentially powerful method for identifying the genes underlying trait variation [e.g., maize (2) and Arabidopsis (3)]. In this study, we used rapid sequencing technology (Illumina next generation DNA sequencing technology) to examine nucleotide diversity in Medicago truncatula, a relative of alfalfa. The patterns of diversity reveal selection acting against most mutations that alter protein sequences, high diversity in genes involved in biotic interactions, and low levels of recombination relative to mutation. Importantly, the sequence information provides a powerful basis for conducting GWAS in M. truncatula.

M. truncatula is a predominantly self-fertilizing plant species that serves as a model for investigating the genetics and genome evolution of legumes as well as cooperation between legumes and the bacteria that fix nitrogen when in root nodules and between plants and beneficial fungi. We sequenced an average of 82 million 90-base paired end reads from each of 26 M. truncatula lines. After applying strict quality filters to the sequence information, we identified more than 3 million genome markers known as SNPs, which revealed the nucleotide diversity of this legume to be slightly greater than the diversity of Arabidopsis thaliana or soybean but lower than the diversity of the highly diverse maize. Moreover, nucleotide diversity was found to be high in classes of genes with well-established roles in defense against pathogens and/or control of differentiation of the symbiotic bacteria involved in nitrogen fixation in the nodules. The high diversity in these gene families may reflect selection for rare genetic variants or reduced selection acting on nonfunctional members of these large gene families.

In multiple species, genomic regions with low levels of recombination harbor less nucleotide variation than those regions with high recombination (4), and we find this pattern in M. truncatula as well (Fig. P1). This phenomenon is consistent with both background selection (i.e., selection against deleterious mutations) and genetic hitchhiking (i.e., linkage to beneficial mutations that are fa-
vored by selection) that reduce genetic diversity at linked sites throughout the genome. Two aspects of our data, an excess of rare nonsynonymous (amino acid changing) relative to synonymous (mutations that do not affect the amino acid sequence of a protein) mutations and the absence of a correlation between gene density and rare variants, suggest that the negative correlation between recombination and nucleotide diversity in *M. truncatula* is primarily because of background selection rather than hitchhiking. Consistent with a limited role for hitchhiking, we found that few regions in the genome exhibited strong signatures of recent selective sweeps.

Linkage disequilibrium (LD) is the nonrandom association between two or more genetic variants that may or may not be in the same chromosomal region. Relative to species that undergo outcrossing, those species that are predominantly self-fertilizing might be expected to have extended LD among polymorphisms (sequence variants) and low rates of effective recombination. Within our sample, however, approximately two-thirds of the SNPs detected were not in complete LD with an adjacent SNP. Moreover, although LD was highly variable (spanning the entire range from absence of to complete LD over distances from 1 to 10 kb), the average LD extended less than 5 kb. At the genome scale, these patterns are very similar to those patterns found in the predominantly self-fertilizing *A. thaliana*. Interestingly, neither of these species shows patterns consistent with the hypothesis that a self-fertilizing species possesses extensive LD, which would hinder adaptive evolution. However, the patterns in both species support the expectation that the evolutionary transition from outcrossing to self-fertilizing will have a much greater effect on recombination than mutation (5).

Based on our analyses and the current costs of whole-genome resequencing, a tagged SNP approach for conducting GWAS in *M. truncatula* does not provide a clear advantage over whole-genome resequencing. In particular, a tagged SNP approach designed to assay all common SNPs detected in our sample would require more than 800,000 tagged SNPs. Moreover, such a strategy would entail substantial bias and impede assaying of low-frequency SNPs. Avoiding these limitations would decrease the probability of identifying potentially misleading synthetic associations (1) and increase the power to correctly identify causal variants and characterize the genetic architecture of complex traits.