Fine-Scale Population Recombination Rates, Hotspots, and Correlates of Recombination in the *Medicago truncatula* Genome

Timothy Paape[1,2,*], Peng Zhou[3], Antoine Branca[4], Roman Briskine[3,5], Nevin Young[3], and Peter Tiffin[1]

[1]Department of Plant Biology, University of Minnesota

[2]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland

[3]Department of Plant Pathology, University of Minnesota

[4]Institute for Biodiversity and Evolution, Westfälische Wilhelms-Universität, Münster, Germany

[5]Department of Computer Science, University of Minnesota

*Corresponding author: E-mail: tim.paape@ieu.uzh.ch.

Accepted: 26 April 2012

Abstract

Recombination rates vary across the genome and in many species show significant relationships with several genomic features, including distance to the centromere, gene density, and GC content. Studies of fine-scale recombination rates have also revealed that in several species, there are recombination hotspots, that is, short regions with recombination rates 10–100 greater than those in surrounding regions. In this study, we analyzed whole-genome resequence data from 26 accessions of the model legume *Medicago truncatula* to gain insight into the genomic features that are related to high- and low-recombination rates and recombination hotspots at 1 kb scales. We found that high-recombination regions (1-kb windows among those in the highest 5% of the distribution) on all three chromosomes were significantly closer to the centromere, had higher gene density, and lower GC content than low-recombination windows. High-recombination windows are also significantly overrepresented among some gene functional categories—most strongly NB–ARC and LRR genes, both of which are important in plant defense against pathogens. Similar to high-recombination windows, recombination hotspots (1-kb windows with significantly higher recombination than the surrounding region) are significantly nearer to the centromere than nonhotspot windows. By contrast, we detected no difference in gene density or GC content between hotspot and nonhotspot windows. Using linear model wavelet analysis to examine the relationship between recombination and genomic features across multiple spatial scales, we find a significant negative correlation with distance to the centromere across scales up to 512 kb, whereas gene density and GC content show significantly positive and negative correlations, respectively, only up to 64 kb. Correlations between recombination and genomic features, particularly gene density and polymorphism, suggest that they are scale dependent and need to be assessed at scales relevant to the evolution of those features.

Key words: recombination, centromere, gene density, GC content, autocorrelation, wavelet analysis.

Introduction

Characterizing genome-wide patterns and rates of recombination are fundamental to understanding how chromosomes evolve, where recombination occurs most frequently, and how genes are distributed. Traditionally, estimating recombination rates was accomplished through genotyping the progeny of experimental crosses (reviewed in Mezard 2006). Cross-based approaches provide direct estimates of recombination but are limited by the number of recombination events that occur in a generation, thereby limiting the scale of resolution to hundreds of kilobase or megabases (Nordborg et al. 2002; Anderson et al. 2004; Gore et al. 2009). Coalescent-based methods allow estimation of the population-scaled recombination rates ($\rho = 4\ N_e r$, where N_e is the effective population size and r is the recombination rate) from genome-wide single-nucleotide polymorphism (SNP) data from a sample of individuals within a species (Fearnhead and Donnelly 2002; McVean et al. 2002; Stumpf and McVean 2003; Fearnhead 2004). These approaches provide recombination estimates at a much finer scale than

map-based estimates and are able to capture the species wide recombination history.

Recombination rates often show broad patterns across genomes, though these patterns differ among species. For example, recombination rates are positively correlated with distance from the centromere in *Drosophila* (Begun and Aquadro 1992), maize (Anderson et al. 2004; Gore et al. 2009; Schnable et al. 2009), wheat (Akhunov et al. 2003), and rice (Wu et al. 2003) but show no correlation with distance from the centromere in *Arabidopsis* (Mezard 2006; Kim et al. 2007) and are positively correlated with distance from the centromere in *Medicago truncatula* at the 100 kb scale (Branca et al. 2011). Recombination rates are also positively correlated with gene density in several species, including humans (Freudenberg et al. 2009), rice (Tian et al. 2009; Flowers et al. 2011), and maize (Anderson et al. 2005; Gore et al. 2009). Although broad patterns can be detected, recombination rates are highly heterogeneous and can vary widely between adjacent 5- and 10-kb regions (Morrell et al. 2006; Buckler and Gore 2007; Kim et al. 2007). As such, patterns at coarse scales may not accurately reflect relationships between recombination and genomic features at finer scales.

Recombination hotspots, defined as short intervals where the local recombination rate greatly exceed recombination in surrounding regions, can only be identified by examining at scales finer than the 50- to 100-kb window scale that is typical of genomic analyses (Drouaud et al. 2006; Kim et al. 2007; Kulathinal et al. 2008) or using genetic maps. Because recombination hotspots may delineate regions that are inherited as linkage blocks, they may provide important insight into the processes that result in the considerable variation in linkage disequilibrium (LD) that is observed in most genomes. Moreover, because a large portion of the recombination that occurs may occur in hotspots (i.e., ~60% of human meiotic events appear to occur within identified hotpots, Coop et al. 2008), uncovering the genomic organization of hotspots may provide insight into the recombination landscape of a genome. Recombination hotspots have been less thoroughly investigated in plant species than in humans, where there are an estimated 25,000 1- to 2-kb hotspot regions (Jeffreys et al. 2001; McVean et al. 2004; Winckler et al. 2005), other mammals (Kauppi et al. 2004; Baudat et al. 2009), *Drosophila* (Stevison and Noor 2009) or yeast (Gerton 2000; Birdsell 2002). Nevertheless, analyses of the progeny of experimental crosses have shown evidence for recombination hotspots in maize (Dooner and Martinez-Ferez 1997; Fu et al. 2001), *Arabidopsis thaliana* (Drouaud et al. 2006), and wheat (Saintenac et al. 2010).

Our goal in this study is to characterize the recombination landscape in the model legume *M. truncatula*. Specifically, using whole-genome resequencing data for three chromosomes from a range-wide collection of 26 individuals, we estimated population-scaled recombination rates at the resolution of 1-kb windows and identified the locations of recombination hotspots at 2-kb resolution with 1-kb overlap. Next, to identify the features of the genome that may explain high recombination, we examine how recombination rates and hotspots vary among chromosome arms, chromosomal location, GC content, gene density, and gene functional categories. The relationship between recombination rates and distance to the centromere differs across species (reviewed in Marais et al. 2001; Nachman 2002; Jensen-Seaman et al. 2004; Drouaud et al. 2006; Mezard 2006; Gaut et al. 2007) with the most common pattern being suppressed recombination immediately near the centromere and a gradient of increasing recombination toward telomeres. GC content is positively correlated with recombination in many organisms (Galtier et al. 2001; Jensen-Seaman et al. 2004; Meunier and Duret 2004; Duret and Arndt 2008), particularly outcrossing species (Marais 2003), but others show no correlation (Drouaud et al. 2006; Mezard 2006). Studies in humans and *Arabidopsis* (Myers et al. 2005; Horton et al. 2012) show that recombination appears highest within intergenic regions, but there is also evidence of elevated recombination rates in tandemly arrayed genic (TAG) regions in plants (Rizzon et al. 2006; Gaut et al. 2007), and positive correlations between recombination and gene density have been reported for rice, maize, and wheat (Liu et al. 2009; Saintenac et al. 2010; Flowers et al. 2011). Finally, because the relationships between recombination estimates and genomic features may be highly dependent upon the scale at which they are examined, we apply wavelet analysis (Spencer et al. 2006; Thurman et al. 2007) to examine how correlations between recombination rates and genomic features vary with spatial scale. We focus our analyses on three of the eight *M. truncatula* chromosomes that have fairly high sequence coverage (Branca et al. 2011), show variation in recombination rates at 100 kb scales, and differ in their gene content and organization (Young et al. 2011).

Materials and Methods

We estimated population-scaled recombination rates $\rho = 4\ N_e r$ (where N_e is the effective population size and r is the recombination rate) and locations of recombination hotspots using sequence data for three chromosomes from a range-wide collection of 26 accessions of the model legume *M. truncatula*. Population-scaled estimates of ρ are affected by the nature of the sample where ours was a range-wide sample and thus may not accurately capture effective recombination within local subpopulations, the effective population size, and actual numbers of recombination events occurring within a region. In brief, the data were obtained using the Illumina sequencing platform to sequence each individual to 15× mean aligned

depth. Paired-end 90 bp reads were aligned to a reference genome sequence (Young et al. 2011), and variants were identified if a base differed from the reference at ≥2 or more uniquely aligned reads and >70% of total reads called the base. Because *M. truncatula* is naturally highly self-fertilizing (selfing rates are estimated to be >98%, Ronfort et al. 2006) and the lines we sequenced were selfed for more than three generations prior to DNA extraction, we did not call any sites as heterozygous (preliminary analyses indicated heterozygous sites were very rare). Prior to estimating recombination rates, we removed all sites segregating more than three bases, had a minor allele frequency <0.1, or were present in fewer than 20 of the 26 sequenced individuals. Further details about these data are in Branca et al. (2011). After applying these filters, there remained 473,502 SNPs across 101,794,000 Kb of the three chromosomes we analyzed.

Estimate of 4Ner (ρ)

To estimate the population-scaled recombination rate, ρ, we used the "interval" program in the LDhat package (McVean et al. 2002; http://www.stats.ox.ac.uk/~mcvean/LDhat). This coalescent-based composite likelihood approximation method can efficiently handle large amounts of data (Auton and McVean 2007), allows direct implementation of fine-scale recombination rate estimates and hotspot detection (McVean et al. 2004; Myers et al. 2005; Winckler et al. 2005), and considers spatial autocorrelation in recombination rate estimates (Smith and Fearnhead 2005). We ran the MCMC algorithm implemented in LDhat interval on 100-kb nonoverlapping sliding windows for 1,000,000 generations sampling every 2,000 generations after a 15,000 generation burn-in. To estimate the recombination rates at 1 kb scales, we estimated ρ for each pair of SNPs and averaged them within each 1-kb window. To evaluate the goodness of fit of the recombination model, a custom likelihood lookup table was created for our system parameterized using 26 genomes considered haploid due to complete homozygosity through selfing and using a mean genome-wide estimate of diversity measured as $\theta_W = 0.006$ bp^{-1} (Branca et al. 2011). We did not include windows with less than five SNPs per window because the accuracy of ρ estimates may be low when estimated using few SNPs. We also excluded estimates from windows with more than 25 SNPs because of the possibility that these highly polymorphic windows reflect errors in SNP variant calling, leaving a remaining 34,356 windows. Because this filtering may affect correlations and include comparisons of 1 kb filtered and unfiltered data set, correlations are presented in supplementary table S1 (Supplementary Material online).

Hotspot Detection

To identify windows with recombination rates significantly higher than mean background rates, that is, recombination hotspots, we used the sequenceLDhot program (Fearnhead 2006; http://www.maths.lancs.ac.uk/~fearnhea/Hotspot/). The estimate of recombination hotspots is a model-based maximum likelihood estimate that compares local recombination rates with surrounding ρ using a null model (H_0: recombination rate in putative hotspot region = background rate) where the background rate (estimated from LDhat) of recombination within a window is equal to the estimated hotspot recombination rate (ρ hat $\hat{\rho}$). Using a likelihood ratio (LR) test, we compare the alternative model (H_A) which states that the recombination rate in a putative hotspot region is 10–100 times the background rate using the LR statistic $\Lambda = (-2[\ln H_0/\ln H_A])$. We ran the program for 500,000 generations over 2-kb sliding windows (1-kb overlap), sampling every 100th generation, and the default setting of seven SNPs to estimate ρ for each window. To evaluate the goodness of fit of the model to the data, we used a likelihood lookup grid that was specified to have an absolute range of r from 0.5 to 40 kb^{-1}, a range that included the range of values estimated from LDhat to estimate mean background rates among windows, we used a LR test to test whether recombination rate in a putative hotspot region is significantly greater than the background rate. The mean background rate was implemented into sequenceLDhot using the previously estimated interval data. We considered windows with LR scores in the highest 5% of all scores (corresponding to LR ≥21, $P = 0.0005$) as hotspots. Under the assumption that hotspots are determined by short chromosomal regions (Fearnhead 2006; Auton and McVean 2007) when adjacent 1-kb windows possessed LR scores ≥21, only the putative hotspot with the highest LR was kept for further analyses (Fearnhead P, personal communication).

Statistical Analyses

We used logistic regression to determine whether genomic features, high-recombination windows, defined as those windows with ρ estimates among the highest 5%, and low-recombination windows, defined as those windows with estimates of ρ below the genome-wide median estimate (fig. 2), were differentially distributed among chromosome arms or were significantly related to GC content, gene density, nucleotide polymorphism (estimated as was estimated as Watterson's diversity estimator, θ_W) or distance to the centromere for each 1-kb window. Logistic regression was conducted using the glm function in R, and the statistical significance of each explanatory variable was evaluated using type III sums of squares (i.e., each factor was evaluated after removing variation attributable to other factors included in the model). For these analyses, we identified GC content as the proportion of 1-kb window with GC, gene density as proportion of coding sequence (coding region only; no introns or untranslated regions) in 1 kb, and

relative distance to the centromere (divided by the length of each chromosome arm) where centromere positions are defined in Young et al. (2011). Similar analyses were conducted to test whether windows identified as recombination hotspots were differentially distributed among chromosome arms or were significantly related to GC content, gene density, θ_W, or distance to the centromere.

To test whether high-recombination windows are nonrandomly distributed on each arm, we calculated the median distance between adjacent high-recombination windows and compared this distance with the median of 100 randomly selected windows of the same number from each arm. This was done similarly for hotspots. Next, to determine whether genes of different function categories were overrepresented in high-recombination regions or hotspots, we used a χ^2 test to determine whether genes that are assigned to different functional categories were nonrandomly distributed between these regions. For these analyses, we considered only gene functional categories for which >30 genes were found in the windows that we had recombination estimates. Assignment of genes to functional categories was based on annotation by the International Medicago Genome Annotation Group (medicago.org/genome/IMGAG/). Similar analyses were used to test whether hotspots were nonrandomly distributed on each chromosome arm or gene functional categories.

We tested for significant linear relationships between ρ and genomic features using Pearson's correlations. Because 1-kb windows are not independent from one another, due to spatial autocorrelation in the data (Hahn 2006), we evaluated the significance of these correlations by comparing the calculated correlations with those from 1,000 permuted data sets in which the linear order of each of the two variables was kept intact (Nordborg et al. 2005).

Wavelet Correlations

We used wavelet analysis as described in Spencer et al. (2006) and Thurman et al. (2007) to examine the relationship between recombination and genomic features (distance to the centromere, GC content, and gene density) at different spatial scales along each of the six chromosome arms. In brief, wavelet analysis creates a series of coefficients from a transformed sequence of observations, such as recombination rate and gene density across 2^n window sizes. The coefficients describe variation at successively increasing scales (i.e., 2 kb up to 512 kb), which can then be implemented into a smoothed linear model analysis. Smoothed correlations are essentially the averages of coefficients between similar window sizes. Although we estimated coefficients for windows from 2 to 2048 or 4096 kb depending on the chromosome arm length, windows larger than 512 kb have few data, and thus were dropped from further analyses. Linear models of smoothed detail coefficients (recombination $\sim$ GC + gene density + distance to centromere − 1) were estimated using modified R scripts from Spencer et al. (2006). The linear model is a multiple linear regression with the intercept forced through the origin where significance of each term is evaluated after accounting for variance attributed to other factors in the model. Prior to analyses, ρ estimates were log transformed to meet the assumption of normally distributed residuals. Because missing data creates excessive gaps along chromosomes, for the wavelet analyses, we did not remove windows with either low or very high SNP density as in previous analyses.

Results

Across the six chromosome arms, we estimated a mean $\rho = 0.0026$ recombination events per base pairs per generation with each chromosome showing considerable heterogeneity in ρ (fig. 1). The distribution of ρ was highly skewed toward low values where nearly half of the windows had estimated recombination rates <0.00093 bp^{-1} (fig. 2) consistent with most recombination events occurring in only a small portion of the entire genome. The ρ value in the upper 5% tail is roughly three to four times the mean and approximately 10 times the median value for each arm (table 1). Recombination rates differed significantly among chromosome arms (analysis of variance, $F = 254.34$, df = 5, $P < 0.0001$) with the recombination rate on the left arm of chromosome 3 (Chr3*L*) significantly higher than recombination rate on any of the other five arms. Logistic regression revealed that high-recombination windows, relative to low-recombination windows, were significantly closer to the centromere, had higher gene density and nucleotide diversity, and lower GC content than low-recombination windows (fig. 3; table 2*A*). On each of the six arms, high-recombination windows more spatially clustered than expected by chance with the median distance between adjacent high-recombination windows smaller than the median distance than each of 100 random samples taken for each chromosome arm.

High-recombination windows were also nonrandomly distributed among functional gene categories with five functional categories significantly overrepresented among the high- compared with low-recombination windows and intergenic regions significantly underrepresented (table 3). The two families with the greatest overrepresentation are NB–ARC's (NB: nucleotide binding; ARC: apoptotic protease-activating factor-1, R protein, and CED-4) and leucine-rich repeats (LRRs), both components of known disease resistance genes (NBS–LRR) in plants (Meyers et al. 1999; DeYoung and Innes 2006). Members of these families are often clustered (Meyers et al. 2003) and are found at particularly high density along the left arm of chromosome 3 (Ameline-Torregrosa et al. 2007; Young et al. 2011) and the right arm of chromosome 5 where clusters of LRR's are present in high-

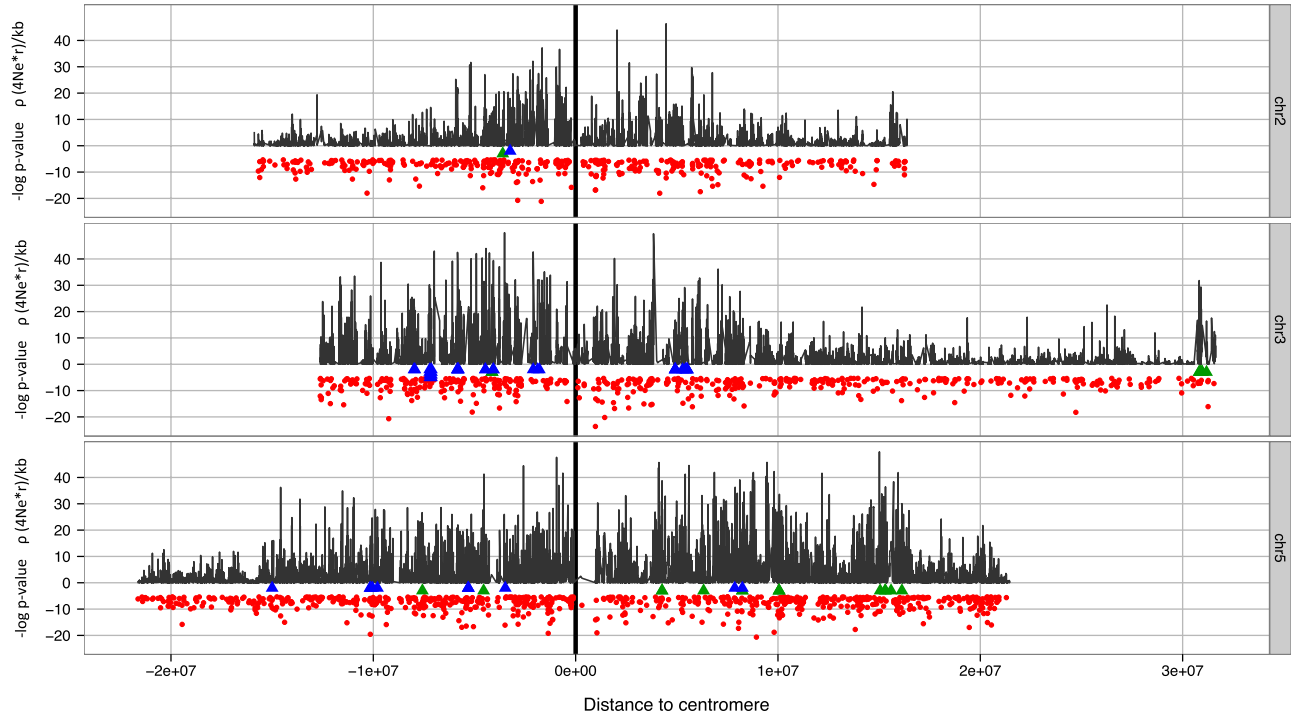

Fig. 1.—Plot of ρ per kilobase along chromosomes 2, 3, and 5 (dark gray) shows variation between the left and right arms (black vertical line is unsequenced centromeric region) as well as distinct variation across chromosomes. Red dots are −log10 of *P*-values calculated from sequenceLDhot LR scores where LR of 21 is the cutoff for significance. Blue triangles represent NB–ARC genes containing windows with significant hotspot LR scores, and green triangles (Chr3*L* and Ch5*R* only) are LRR's windows with significant LR scores.

recombination regions (fig. 1). These are the two arms with highest recombination rates and greatest variance in ρ (table 1). We also found that intergenic regions were significantly overrepresented among low-recombination windows.

Recombination Hotspots

Using sequenceLDHot (Fearnhead 2006), we identified 1902 windows with as recombination hotspots as significant (i.e., the recombination rate is 10–100 times greater than the estimates of recombination in region). Note that recombination hotspots and high-recombination windows are not synonymous—recombination hotspots are detected relative to background rate of recombination, whereas high-recombination windows are those with high rates relative to the surrounding genomic background rate. Of these 1902 recombination hotspot windows, 466 were adjacent to other hotspot windows. For these adjacent windows, we kept only the window with the highest statistical support for subsequent analyses, leaving 1669 hotspot windows (fig. 1). The windows identified as recombination hotspots are clearly distinct from the high-recombination windows, with only 94 windows identified as both high-recombination windows and recombination hotspots.

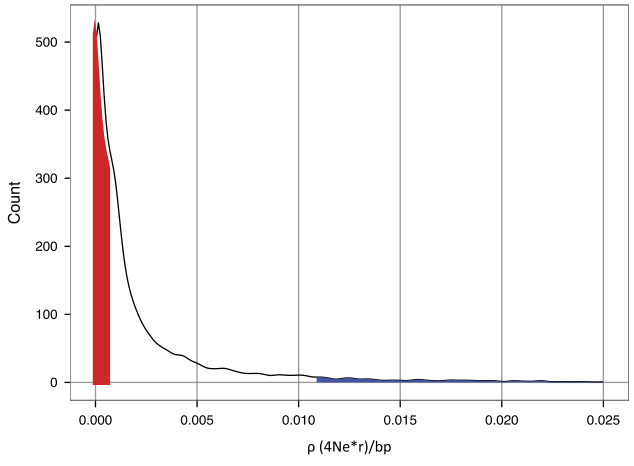

Fig. 2.—Distribution of ρ across chromosomes 2, 3, and 5. Red is 0 median (lower 50% tail), and blue is upper 5% tail.

Unlike high-recombination regions, we detected no significant relationship between hotspots windows and GC content ($P = 0.49$) or gene content ($P = 0.95$; table 2). By contrast, hotspot windows were significantly closer to the centromere than nonhotspot windows ($P < 0.001$), and hotspot windows had significantly greater diversity than nonhotspots for all chromosome arms. Both these patterns are similar to what was seen with high- compared with low-recombination windows. We also find a significant effect of polymorphism with hotpots using logistic regression (table 2) where hotspots had significantly greater diversity than

Table 1
Mean, Standard Deviation (SD), and Quantile Values for ρ on Each Chromosome Arm and the Number of (*N*) 1-kb Windows and Total Length in Base Pairs

Chromosome arm	Mean	SD	Median	95% Cutoff	Maximum	*N* Windows	Length (bp)
Chr2*L*	0.0019	0.0036	0.0008	0.0088	0.0372	4,477	15,870,317
Chr2*R*	0.0016	0.0030	0.0007	0.0064	0.0463	4,350	16,118,084
Chr3*L*	0.0042	0.0064	0.0017	0.0176	0.0500	3,494	12,626,695
Chr3*R*	0.0020	0.0037	0.0008	0.0085	0.0496	6,953	31,531,826
Chr5*L*	0.0024	0.0040	0.0010	0.0103	0.0476	7,761	21,581,903
Chr5*R*	0.0035	0.0057	0.0012	0.0156	0.0497	7,330	21,200,016

nonhotspots for all chromosome arms. We do however see significantly higher mean GC content for hotspots than for high ρ windows for Chr3*L* (fig. 3). Also, unlike high-ρ windows, hotspots show less evidence of being spatially clustered; the median distance between adjacent hotspot windows was smaller than 95% of resampled windows only for Chr2*R* and Chr5*R*. This however could be an artifact of the LR cutoff used to define hotspots. When the cutoff is reduced to LR = 12 (P = 0.0005), a value used in several human hotspot studies (Fearnhead 2006) and median distances between hotspots for all chromosome arms except Chr3*L* are always less than 100 resampled distances. Also, unlike the high-recombination windows, hotspot windows are not significantly overrepresented in particular regions and in fact are significantly underrepresented among NB–ARC's genes (χ^2 = 4.21, P = 0.042) and no other functional gene categories show significant overrepresentation (P-values between 0.11 and 0.86), for which high-recombination windows were the most significantly overrepresented.

Correlates of Recombination and Wavelet Analysis

Pairwise correlation coefficients calculated using data from each 1-kb window reveal that for each of the six chromosome arms, ρ was negatively correlated with the distance to centromere, positively correlated with gene density, negatively correlated with GC content, and positively correlated with nucleotide polymorphism (table 4, supplementary

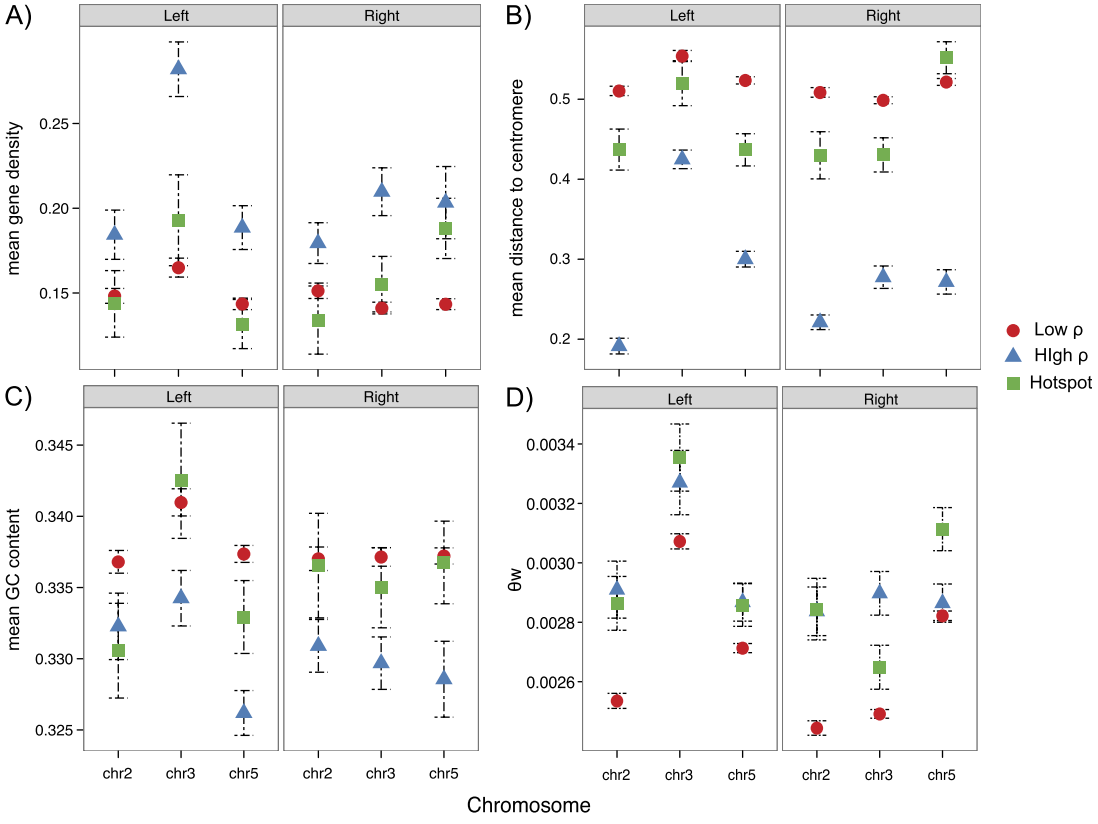

Fig. 3.—Mean comparisons of gene density (*a*), distance to centromere (*b*), GC content (*c*), and θ_w (*d*) between regions of high ρ (upper 5% tail = blue), lower 50% tail (red), and estimated hotspots (green). Error bars represent 95% standard errors.

Table 2
Results from Logistic Regression Testing Whether (*A*) Low- and High-ρ Windows and (*B*) Hotspot Versus Nonhotspot Windows Differ between Chromosome Arms and Distance to Centromere, GC Content, and Gene Density

Genomic Category	Df	χ^2	*P*-Value
(A) Low versus high ρ			
GC content	1	120.8	>0.001
Gene density	1	227.8	>0.001
Distance to centromere	1	1219.4	>0.001
θ_w	1	153.5	>0.001
Chromosome arm	5	112.6	>0.001
Residuals	19,369		
(B) Hotspot versus nonhotspot			
GC content	1	7.14	0.49
Gene density	1	0.47	0.95
Distance to centromere	1	0.004	0.007
θ_w	1	35.55	>0.001
Chromosome arm	5	3.99	0.55
Residuals	34,356		

table S1, Supplementary Material online). Although most of these correlations with recombination rates are statistically significant, the magnitude of the correlations is generally small, and correlation coefficients are greater than 0.1 (or $\leq$0.1) for distance to the centromere only. Although gene density appears to be generally weakly (though significantly) correlated with recombination in Pearson correlations, with Chr3*L* showing the strongest positive correlation with gene density (supplementary table S1, Supplementary Material online), this is supported by the results of figure 3 where this arm had the highest mean gene density. Weak negative correlations are occasionally seen when using the unfiltered data set but not significantly so (supplementary table S1, Supplementary Material online). Most importantly, although we found previously in Branca et al. (2011) and in the current analysis, correlations between ρ and gene density are negative for all arms (supplementary table S2, Supplementary Material online). However, because of spatial autocorrelation of variables along the genome, 1-kb windows are not independent of one another, and therefore, Pearson's correlation coefficients based on the number of 1-kb windows and their statistical significance should be viewed with caution (Hahn 2006).

Smoothed correlations from wavelet analysis provide insight into correlations across multiple spatial scales while also removing the nonindependence between variables. Because distance to the centromere is included in these multiple linear regressions, the effects of spatial autocorrelation are removed before examining the relationships between recombination and gene density and GC content. With the exception of two chromosome arms (Chr2*R* and Chr3*L*), the smoothed linear wavelet models show significant negative correlations between recombination and distance from centromeres at scales up to 512 kb (fig. 4). For Chr3*L*, negative correlations between recombination and distance to centromere are found up to 128 kb but disappear at greater scales reflecting different patterns across the short and long arms of this chromosome. Significant positive correlations between recombination and gene density for wavelet models are found only on Chr3*L* and Chr5*R* but only at short scales (2–16 kb); these are the two chromosome arms with the highest mean ρ. We find significant negative correlations between recombination and GC content at small to intermediate scales (up to 64 kb) on all arms except Chr3*L* (fig. 4). The loss of significant correlations at larger scales is likely due to smaller sample sizes and thus less statistical power, and typically, a reduction in magnitude at large scales is likely due to averaging across heterogenous regions contained in the larger windows.

Table 3
Functional Categories Overrepresented in High-Recombination Windows

	Low ρ		High ρ		
IMGAG Function	**Expected**	**Actual**	**Expected**	**Actual**	***P*-Value**
Intergenic	6,643	6,841	664	588	***0.011***
NB–ARC	209	122	21	49	***<0.0001***
Protein kinase	224	204	22	24	0.60
LRR	101	74	10	27	***0.0007***
Zinc finger	220	238	22	26	0.77
Cyclin-like F-box	177	181	18	26	0.26
2OG-Fe(II) oxygenase	49	32	5	12	***0.018***
Glycoside hydrolase	113	104	11	10	0.93
Pentatricopeptide repeat	81	99	8	10	0.98
UDP-glucosyl-transferase	46	34	5	10	0.067
Cytochrome P450	83	51	8	9	0.26
Cellulose synthase	18	10	2	7	***0.023***
Lipolytic enzyme	34	17	3	7	***0.042***
Pectinesterase inhibitor	18	11	2	7	***0.030***

NOTE.—Functional categories are ordered according to number found in high-recombination (upper 95% tail) regions. Bold are genes that are significantly overrepresented in upper 5% tail.

Table 4
Full Correlation of 1-kb Windows between Recombination Rate (ρ) and Gene Density, GC Content, and Distance to the Centromere

	Chromosome Arm					
	2L	**2R**	**3L**	**3R**	**5L**	**5R**
Gene density	**0.045**	**0.043**	**0.126**	**0.078**	**0.05**	**0.05**
P-value	*0.015*	*0.025*	*0.001*	*<0.001*	*0.008*	*0.005*
GC content	**−0.046**	**−0.08**	**−0.046**	**−0.05**	**−0.08**	**−0.08**
P-value	0.06	*0.031*	*0.017*	*0.009*	*0.005*	*0.004*
Distance to the centromere	**−0.28**	**−0.15**	**−0.13**	**−0.21**	**−0.22**	**−0.145**
P-value	*0.003*	0.339	0.31	*0.038*	*0.007*	0.259
θ_w	**0.063**	**0.071**	**0.032**	**0.094**	**0.05**	**0.02**
P-value	*0.04*	*0.03*	0.096	*0.018*	0.06	0.17

NOTE.—Pearson's correlation coefficients are in bold, *P*-values < 0.05 are in italics.

Discussion

In a previous analysis of recombination in this same sample of *M. truncatula*, we found LD decay decreases within 5 kb on average across the genome (Branca et al. 2011). This rate of decays suggests that 100 kb scales, which we used previously when characterizing nucleotide diversity and recombination may be too coarse to characterize the genomic features that shape recombination. For this reason, we examined recombination rates at the 1 kb scale along three of the eight *M. truncatula* chromosomes. A comparison of high-recombination windows (those in highest 5% of the distribution) with low-recombination windows (those in the lowest 50% of the distribution) revealed that high-recombination windows are more likely to be closer to the centromere, have higher GC content, and found in regions of higher gene density. High-recombination windows also are significantly overrepresented among a subset of common gene families, particularly NB–ARC's and LRR's. Finally, we found stronger relationships between recombination rates and genomic features at smaller than larger scales of resolution—this is likely due to larger windows averaging across smaller scales at which recombination rates and GC, gene content, and diversity vary.

The genomic features associated with high-recombination regions in *M. truncatula* appear to differ from those found in several other organisms. For example, in *Medicago*, we find higher recombination rates nearer centromeric regions which is opposite to what has been shown for every nonplant system, including *Caenorhabditis elegans*, *Drosophila* (Marais et al. 2001), mice, rats (Nachman 2002, Jensen-Seaman et al. 2004), and humans (Freudenberg et al. 2009), and most plants (Mezard 2006; Gaut et al. 2007). This negative gradient in recombination is in stark contrast to wheat and maize, both of which show clear patterns of recombination increasing with relative distance from the centromere (Lukaszewski and Curtis 1993; Akhunov et al. 2003; Gore et al. 2009). By contrast, *A. thaliana* shows highly variable levels of recombination along each chromosome (Kim et al. 2007; Horton et al. 2012) with no significant centromere–telomere gradient (Drouaud et al. 2006; Mezard 2006). In wheat, Lukaszewski and Curtis (1993) also found a stronger gradient of increasing recombination along short chromosome arms relative to long ones.

Our results also reveal that gene density is greater in high-recombination than low-recombination regions of the genome. This pattern is particularly strong on Chr3*L*, a chromosome arm that includes several NB–ARCs and LRRs, as well as shorter nodule cysteine-rich peptides and defensin-like proteins (Young et al. 2011), many of which are found as highly TAG clusters (Graham et al. 2004; Silverstein et al. 2005; Ameline-Torregrosa et al. 2007). We find resistance-related genes in general to be significantly overrepresented in *M. truncatula* for high-recombination regions consistent with what has been shown for human immune loci (Jeffreys et al. 2001; McVean et al. 2004; Winckler et al. 2005) and with recent findings for disease resistance genes showing high recombination in *A. thaliana* (Horton et al. 2012). Among the 60 annotated NB–ARC's on Chr3*L* in this data set, 18 are found in the upper 5% ρ tail, and we are therefore not surprised to find high recombination and gene density to be correlated when these genes are present. High-recombination rates in TAG regions are also found in *A. thaliana* (Zhang and Gaut 2003), rice (Rizzon et al. 2006), and wheat (Akhunov et al. 2003). However, the overrepresentation of high-ρ windows ($\rho = N_e r$) among members of NB–ARCs and LRRs may be due to elevated N_e rather than higher r; members of these gene families having significantly elevated levels nucleotide diversity *M. truncatula* (Branca et al. 2011).

In *M. truncatula*, we find a weak correlation between recombination and GC content at the 1 kb scale in pairwise correlations and negative but decreasing in magnitude and significance at increasing scales. This is similiar to the negative correlation between recombination and GC content found in *A. thaliana* by Drouaud et al. (2006); yet, when we use means over 100 kb scales in pairwise correlations (supplementary table S2, Supplementary Material online), the correlations are generally positive. Positive correlations are found between recombination and GC for most outcrossing animals (Marais et al. 2001; Jensen-Seaman et al. 2004; Duret and Galtier 2009) and recent comparative studies in selfing and outcrossing plants confirm that reduced GC bias is predicted by mating system (Muyle et al. 2011; Qiu et al. 2011). Drouaud et al. (2006) indicated that GC and recombination may not be correlated with one another in highly selfing species because selfers do not form heteroduplex DNA during recombination because of high homozygosity (Marais et al. 2004), but the lack of heteroduplex formation does not explain a significant negative correlation between ρ and GC. Because we see a consistent negative correlation between GC content and using regression and linear models for all chromosomes, this appears to

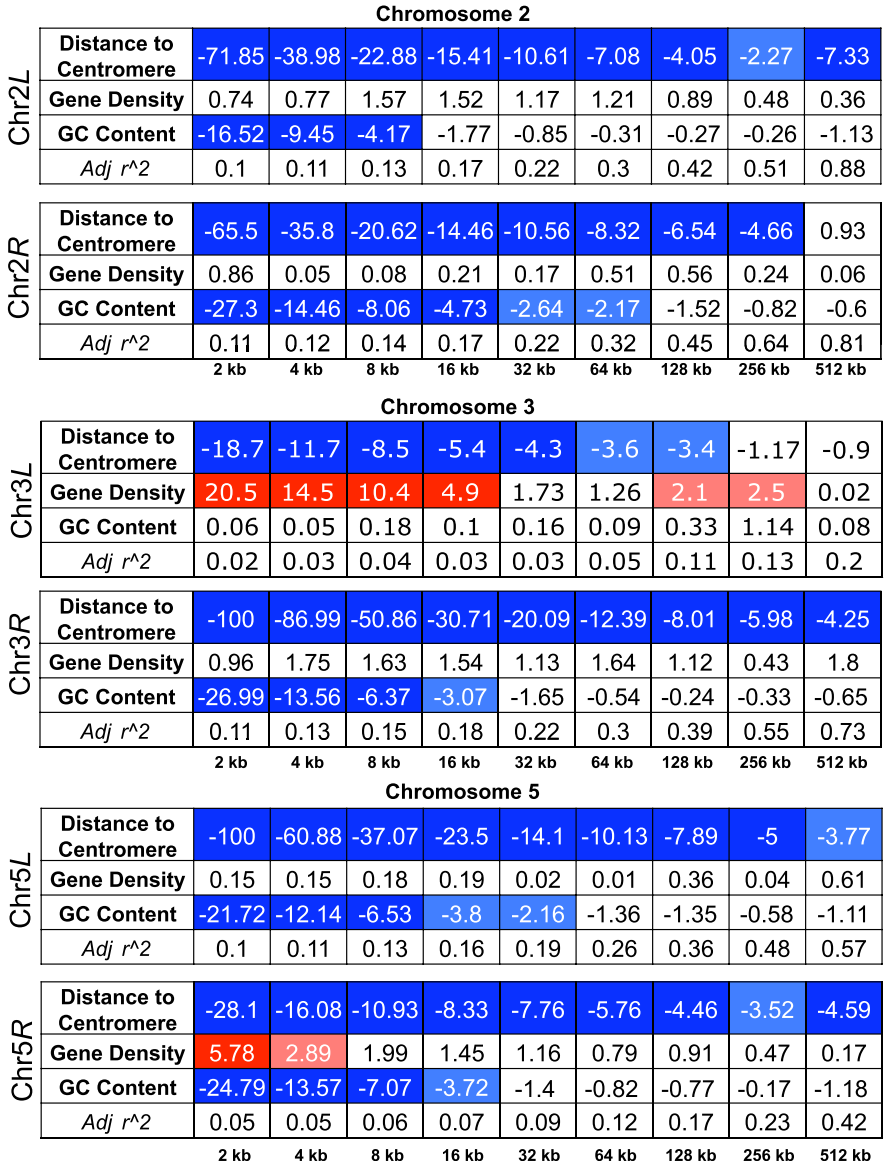

Fig. 4.—Smoothed wavelet correlations between recombination (ρ), relative distance to the centromere, gene density, and GC content. Red indicates positive linear relationship and blue indicates negative linear relationship (using *t*-tests to determine $-\log_{10}$ of *P*-values). Magnitude of the color is proportional to the level of significance.

be a genome-wide pattern in *M. truncatula* but more pronounced at scales between 2 and 64 kb (fig. 4) suggesting potential effects of gene density covariation.

The fine-scale (1 kb) analyses of recombination rates reported here are largely consistent, although not always identical with previous analyses that examine recombination in 100-kb windows reported in Branca et al. (2011). There are however two differences; the previous analyses detected a significant genome-wide negative correlation between population-scaled recombination rate and gene density, whereas at the 1 kb scale, we detect slight positive correlations, and the correlation between recombination and diversity is weaker at the 1 kb than 100 kb scales (table 4; supplementary table S2, Supplementary Material online). There are two aspects of the data that may be responsible for the different relationships seen at different scales of analyses. First, for the 1-kb analyses, we excluded low-diversity windows because of the expectation that these windows would harbor little information for accurately estimating recombination rates. In fact, correlations calculated using the full data set are generally weaker and less often significantly different from zero than those calculated using the filtered data (supplementary table S1, Supplementary Material online). The second possible reason is that the larger windows average across heterogenous genome regions; this certainly appears to contribute to the differences given that even at

100-kb resolution, the correlation between recombination and gene density differs among chromosome arms (positive correlations on some arms and negative on others). Regardless of the reasons, it is clearly important to consider scale when most genome-wide studies of recombination in plants have been observed at 100 kb or greater (reviewed in Flowers et al. 2011).

Recombination Hotspots

Windows with high levels of r are clearly distinct form those identified as recombination hotspots, which are defined as having recomibination rates that are 10–100 times greater than surrounding regions. Unlike high-ρ windows, hotspot windows do not differ significantly from nonhotspot windows in GC content or gene density. We do however find that hotspots are significantly closer to the chromosome and have significantly greater nucleotide diversity than nonhotspot windows. Higher nucleotide diversity in hotspot than nonhotspot windows has also been reported for human data (Spencer et al. 2006). This relationship between hotspots and diversity indicate recombination hotpsots may be important in shaping diversity even if they are not stable enough to shape interspecific divergence (Winckler et al. 2005; Spencer et al. 2006; Baudat et al. 2009).

Wavelet Analyses

Wavelet analyses provide an analytical approach to examine relationships between a response variable (ρ) and explanatory variables (distance to centromere, GC content, and gene density) vary with different scales of analyses. The wavelet analyses we implemented, the same as applied by Spencer et al. (2006) to examine recombination in the human genome, uses linear regression that also removes variation due to other explanatory factors in the analyses. In other words, it looks at the relationships between recombination and an explanatory variable after accounting for variation due to other variables in the model. As such, wavelet analyses may reveal scale-dependent patterns that could be missed when conducting analyses with fixed-sized windows. The results from the wavelet analyses indicate that aside from the significant negative correlation between recombination and distance to the centromere across all three chromosomes at scales up to 512 kb, other correlates of recombination show the significance of such correlations disappear around 64 kb or smaller (fig. 4)—this change in statistical significance appears to be due to both loss of statistical power and weaker correlations. In no cases, do we find that correlations switch from positive to negative or vice versa, only in magnitude. The other interesting result from the wavelet analyses is that because it incorporates a multiple linear regression to identify genomic features that are related to recombination rates, it removes colinearity among potential explanatory variables. Removing this colinearity reveals that gene density has only a very weak effect on recombination rate after one accounts for GC content and distance from centromere (supplementary table S1, Supplementary Material online).

Conclusion

Population-scaled recombination rates estimated at 1 kb in *M. truncatula* appear correlated with several genomic factors, including distance to the centromere, gene density, and gene organization. Recombination hotspot regions are also consistently associated with higher diversity across all three chromosomes, a finding that has rarely been discussed for plants. Further insight into recombination rates and hotspots including the genomic features that shape recombination rates could be gained both by examining recombination rates and LD within local subpopulations and combining sequence-based physical maps with high-resolution genetic maps. Most importantly, we find that when analyzing correlated genomic features, scales at which these variables are estimated can produce different results—an important consideration when interpreting results where data are only available at a single or very broad scale.

Supplementary Material

Supplementary tables S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Paul Fearnhead, Gilean McVean, and Adam Auton for useful technical advice and recommendations regarding software and data input. We also thank Kevin Silverstein and Peter Morrell for useful discussions and Brendan Epstein for assistance with script adjustments. This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute and was funded by National Science Foundation Grant 0820005.

Literature Cited

Akhunov ED, et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. 13:753–763.

Ameline-Torregrosa C, et al. 2007. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiol. 146:5–21.

Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS. 2005. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. Genome Res. 16:115–122.

Anderson LK, et al. 2004. Integrating genetic linkage maps with pachytene chromosome structure in maize. Genetics 166:1923–1933.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17:1219–1227.

Baudat F, et al. 2009. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836–840.
Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. Nature 356:519–520.
Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol. 19:1181–1197.
Branca A, et al. 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc Natl Acad Sci U S A. 108:E864–E870.
Buckler E, Gore M. 2007. An Arabidopsis haplotype map takes root. Nat Genet. 39:1056–1057.
Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319:1395–1398.
DeYoung BJ, Innes RW. 2006. Plant NBS-LRR proteins in pathogen sensing and host defense. Nat. Immunol. 7:1243–1249.
Dooner HK, Martinez-Ferez IM. 1997. Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell 9:1633–1646.
Drouaud J, et al. 2006. Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots." Genome Res. 16:106.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Ann Rev Genomics Hum Genet. 10:285–311.
Fearnhead P. 2004. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics 167:2067–2081.
Fearnhead P. 2006. SequenceLDhot: detecting recombination hotspots. Bioinformatics 22:3061–3066.
Fearnhead P, Donnelly P. 2002. Approximate likelihood methods for estimating local recombination rates. J R Stat Soc B. 64:657–680.
Flowers JM, et al. 2011. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. Mol Biol Evol. 29:675–687.
Freudenberg J, Wang M, Yang Y, Li W. 2009. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. BMC Bioinformatics 10:S66.
Fu H, et al. 2001. The highly recombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome. Proc Natl Acad Sci U S A. 98:8903–8908.
Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.
Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet. 8:77–84.
Gerton JL. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A. 97:11383–11390.
Gore MA, et al. 2009. A first-generation haplotype map of maize. Science 326:1115–1117.
Graham MA, Silverstein KA, Cannon SB, VandenBosch KA. 2004. Computational identification and characterization of novel genes from legumes. Plant Phys. 135:1179–1197.
Hahn MW. 2006. Accurate inference and estimation in population genomics. Mol Biol Evol. 23:911–918.
Horton MW, et al. 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet. 44:212–216.
Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. 29:217–222.
Jensen-Seaman MI, et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14:528–538.
Kauppi L, Jeffreys AJ, Keeney S. 2004. Where the crossovers are: recombination distributions in mammals. Nat Rev Genet. 5:413–424.
Kim S, et al. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 39:1151–1155.
Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. 2008. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proc Natl Acad Sci U S A. 105:10051–10056.
Liu SZ, et al. 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 5:e1000733.
Lukaszewski AJ, Curtis CA. 1993. Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. Theor Appl Genet. 86:121–127.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. 5:R45.
Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A. 98:5688–5692.
McVean GA, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241.
McVean G, et al. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.
Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 21:984–990.
Meyers BC, et al. 1999. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J. 20:317–332.
Meyers BC, et al. 2003. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15:809–834.
Mezard C. 2006. Meiotic recombination hot spots in plants. Biochem Soc Trans. 34:531–534.
Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics 173:1705–1723.
Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glemin S. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). Mol Biol Evol. 28:2695–2706.
Myers S, et al. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.
Nachman MW. 2002. Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev. 12:657–663.
Nordborg M, et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 30:190–193.
Nordborg M, et al. 2005. Patterns of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196.
Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. 2011. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. Genome Biol Evol. 3:868–880.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and Rice. PLoS Comput Biol. 2:e115.
Ronfort J, et al. 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.
Saintenac C, et al. 2010. Variation in crossover rates across a 3-Mb contig of bread wheat (*Triticum aestivum*) reveals the presence of a meiotic recombination hotspot. Chromosoma 120:185–198.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115.
Silverstein KA, Graham MA, Paape TD, VandenBosch KA. 2005. Genome organization of more than 300 defensin-like genes in Arabidopsis. Plant Phys. 138:600–610.
Smith NGC, Fearnhead P. 2005. A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. Genetics 171:2051–2062.
Spencer CCA, et al. 2006. The influence of recombination on human genetic diversity. PLoS Genet. 2:e148.
Stevison LS, Noor MAF. 2009. Recombination rates in Drosophila. Encyclopedia of life sciences. Chichester (UK): John Wiley & Sons, Ltd.
Stumpf MPH, McVean GAT. 2003. Estimating recombination rates from population-genetic data. Nat Rev Genet. 4:959–968.
Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. 2007. Identification of higher-order functional domains in the human ENCODE regions. Genome Res. 17:917–927.
Tian Z, et al. 2009. Does genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. 19:2221–2230.
Winckler W, et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111.
Wu J, et al. 2003. Physical maps and recombination frequency of six rice chromosomes. Plant J. 36:720–730.
Young N, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524.
Zhang L, Gaut BS. 2003. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? Genome Res. 13:2533–2540.

Associate editor: Michael Purugganan