MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 1162–1173 doi: 10.1111/j.1365-294X.2010.04559.x

Fine scale genetic structure in the wild ancestor of maize (*Zea mays* ssp. *parviglumis*)

JOOST VAN HEERWAARDEN,* JEFFREY ROSS-IBARRA,* JOHN DOEBLEY,† JEFFREY C. GLAUBITZ,† JOSE DE JESÚS SÁNCHEZ GONZÁLEZ,‡ BRANDON S. GAUT§ and LUIS E. EGUIARTE¶

**Department of Plant Sciences, University of California, Davis, CA 95616, USA, †Department of Genetics, University of Wisconsin, Madison, WI 53706, USA, ‡Centro Universitario de Ciencias Biológicas y Agropecuarias, Universidad de Guadalajara, Zapopan, Jalisco CP45110, Mexico, §Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA, ¶Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, CU, AP 70-275 Coyoacán, 04510 México, DF, México*

Abstract

Analysis of fine scale genetic structure in continuous populations of outcrossing plant species has traditionally been limited by the availability of sufficient markers. We used a set of 468 SNPs to characterize fine-scale genetic structure within and between two dense stands of the wild ancestor of maize, teosinte (*Zea mays* ssp. *parviglumis*). Our analyses confirmed that teosinte is highly outcrossing and showed little population structure over short distances. We found that the two populations were clearly genetically differentiated, although the actual level of differentiation was low. Spatial autocorrelation of relatedness was observed within both sites but was somewhat stronger in one of the populations. Using principal component analysis, we found evidence for significant local differentiation in the population with stronger spatial autocorrelation. This differentiation was associated with pronounced shifts in the first two principal components along the field. These shifts corresponded to changes in allele frequencies, potentially due to local topographical features. There was little evidence for selection at individual loci as a contributing factor to differentiation. Our results demonstrate that significant local differentiation may, but need not, co-occur with spatial autocorrelation of relatedness. The present study represents one of the most detailed analyses of local genetic structure to date and provides a benchmark for future studies dealing with fine scale patterns of genetic diversity in natural plant populations.

Keywords: crop wild relatives, fine scale genetic structure, gene flow, spatial autocorrelation, *zea mays* ssp. *parviglumis*,

Received 11 August 2009; revision received 22 December 2009; accepted 12 January 2010

Introduction

One of the main aims of ecological genetics is to describe and understand the spatial distribution of genetic variation. The organization of genetic diversity within species reflects geographical and demographic factors that affect patterns of mating. Deviation from genome-wide patterns of genetic structure may be informative of selection at individual loci (Lewontin & Krakauer 1973; Beaumont 2005). Genetic structure within continuous populations is of particular interest, as it may be informative of dispersal distances (Hardy & Vekemans 1999) or signal the existence of environmental features that affect gene flow across the landscape (e.g. Piertney *et al.* 1998).

Theoretical work on continuously distributed species dates back to Wright (1943) and Malécot (1948), both of whom addressed the patterns of relatedness in populations that are characterized by decreased mating probability with distance. Subsequent simulation studies have explored the spatial distribution of allelic and genotypic

Correspondence: Jeffrey Ross-Ibarra,
E-mail: rossibarra@ucdavis.edu

© 2010 Blackwell Publishing Ltd

frequencies in two dimensional habitats, both under simple isolation by distance (Rohlf & Schnell 1971; Epperson 1995), and under asymmetrical migration and selection (Sokal *et al.* 1989a). These studies have shown that when mating frequency decreases with distance, single locus genotypes will be correlated in space, but localized spatial patterns shared among unlinked loci are not expected. When directional gene flow occurs, however, localized genetic differentiation of gene frequencies at multiple loci will arise.

Two nonexclusive types of spatial genetic structure can thus be distinguished. The first, which we will refer to as spatial autocorrelation of relatedness, exists when individuals closer together spatially are also more closely related genetically. Spatial autocorrelation of relatedness may arise under many circumstances where mating or dispersal is restricted in space (Slatkin & Arter 1991) but in its simplest form occurs when the dispersal probability of gametes or offspring decreases as a function of distance (Wright 1943). The second type of spatial structure, here termed differentiation, occurs whenever individuals grouped by genetic similarity form distinct geographical or spatial patterns. This form of structure is only expected when underlying biological or environmental factors are spatially nonrandom (e.g. Sokal *et al.* 1989a).

Although both forms of genetic structure have been frequently observed at the landscape level in both plants (Bockelmann *et al.* 2003; Vigouroux *et al.* 2008) and other organisms (Menozzi *et al.* 1978; Barbujani 1987; Sokal *et al.* 1989b; Piertney *et al.* 1998), differentiation has been rarely described within individual populations. The large geographic areas covered by landscape genetic studies (Manel *et al.* 2003) make the detection of differentiation likely, either due to discontinuities in sampling (Handley *et al.* 2007) or because of the presence of strong historical and geographical barriers to gene flow (Castric *et al.* 2001). These factors are expected to be less prominent at small spatial scales, such as within stands of outcrossing plants, making the detection of differentiation in such species challenging. Indeed, most fine-scale genetic studies do not describe detailed patterns of differentiation and have been generally limited to analysis of spatial autocorrelation of relatedness (Loiselle *et al.* 1995; Hardy & Vekemans 1999; Smouse & Peakall 1999; Hardy 2003; Hardy *et al.* 2004; see Jones *et al.* 2007 for an exception). But while spatial autocorrelation analysis is a sensitive method for detecting spatial genetic structure (Sokal & Oden 1978; Epperson & Li 1996), without explicit spatial description of differentiation it cannot distinguish among different explanations for the data (Slatkin & Arter 1991). Observation of decreasing relatedness with distance may, for example, simply reflect symmetrical decay in mating probability, but could also be caused by differentiation due to factors such as kin-structured dispersal or regeneration (Knowles *et al.* 1992; Ingvarsson & Giles 1999), local selection, or asymmetrical patterns of gene flow (Sokal *et al.* 1989a).

Joint analysis of spatial autocorrelation of relatedness and patterns of differentiation poses at least two significant challenges. First, sampling of individuals must be dense and uniform to avoid spurious identification of differentiation due to sampling (Serre & Paabo 2004; Handley *et al.* 2007). Second, the number of available markers for most wild species is limited, resulting in a lack of power to detect differentiation at fine spatial scales (Kalisz *et al.* 2001; Hardy *et al.* 2004). While dense sampling is common in many studies of fine-scale genetic structure in plants, a lack of genetic markers is likely the main reason that so few empirical studies on fine-scale genetic structure have been successful at describing the spatial patterns of localized differentiation.

In this paper we exploit the availability of genome-wide SNP markers to study fine-scale genetic structure in the wild ancestor of cultivated maize, annual teosinte (*Zea mays* ssp. *parviglumis*, hereafter teosinte). Like maize, teosinte is an outbreeding annual grass. It is common throughout south-central Mexico, where it is found growing in large, dense populations. Several authors have investigated patterns of genetic structure in teosinte (Doebley *et al.* 1984; Fukunaga *et al.* 2005; Moeller *et al.* 2007), and while these studies reveal strong population structure at the landscape scale, little is known regarding fine-scale genetic structure. The combination of high diversity and low population differentiation observed in teosinte from the Balsas region of Mexico (Moeller *et al.* 2007) makes these populations of particular interest for studying continuous genetic variation.

We sampled a total of 964 seeds from individually mapped teosinte plants from two populations situated within contiguous grassland habitat in the Balsas region of Mexico. Seeds were sampled uniformly at ~5 m intervals to achieve fine-scale spatial coverage within both sites. Each individual was genotyped using 468 SNP markers, selected from both random and candidate genes (Weber *et al.* 2007, 2008). Our aim was to elucidate spatial patterns of genetic structure at the limits of high gene flow. We specifically addressed the occurrence of differentiation, in combination with spatial autocorrelation of relatedness, by studying spatial trends using principal component analysis and relating them to allele frequency differences. We evaluated the possible roles of kin structure, natural selection and topographical discontinuities as potential factors that could explain the observed genetic structure. Our

© 2010 Blackwell Publishing Ltd

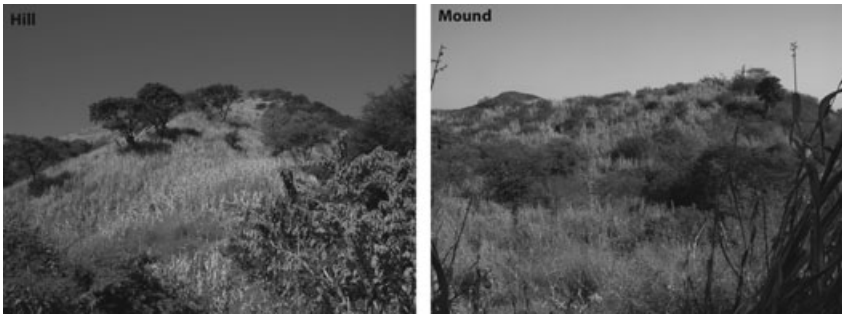

Fig. 1 Images showing the Hill and the Mound sites.

combination of dense, uniform sampling and a large number of markers allowed us to analyse genetic structure in unprecedented detail, making it one of the first empirical studies to describe the precise patterns of spatial genetic differentiation within individual outcrossing populations.

Methods

Plant materials and sampling

Single seeds of individual teosinte plants were collected from two sites in the south of the Mexican state of Mexico (Fig. 1). The two sites—Hill (18.583°N, 100.356°W) and Mound (18.640°N, 100.358°W)—are located 6.3 km apart within large, semi-contiguous stands of teosinte in a mesquite grassland. Seeds were collected from the Hill site on November 30, 2005 from 389 individual teosinte plants in a very dense stand growing on a steep hillside. The eight seed collectors formed a line at the base of the slope roughly parallel to the contour, moving upslope and stopping approximately every five meters to collect seed from a plant. The two seed collectors at the ends of the line took GPS readings every fifth plant; GPS coordinates of other sampled individuals were calculated via linear interpolation of these outside readings. Seeds from 575 plants were collected the following day from the Mound site using a modification of the same collection strategy in which two transects were laid from a central location in the population. An additional 2-m grid was used to finish collection in the northwestern extreme of the Mound population. Field elevation profiles for both sites were inferred by submitting a square grid of spatial coordinates surrounding each site to the NASA SRTM3 elevation database, as implemented on the GPS visualizer website (http://www.gpsvisualizer.com/).

Genotyping

We selected SNPs that had been successfully genotyped (less than 10% missing data) in maize and teosinte (Briggs *et al.* 2007; Weber *et al.* 2007, 2008, unpublished data). From these, we preferentially chose SNPs from alignments that contained multiple scorable SNPs. These SNPs were selected from alignments of either (1) genes with known biological functions (henceforth 'candidate genes'; Weber *et al.* 2007, 2008) or (2) low-copy EST sequences chosen at random from ~10 000 maize ESTs in the MMP-DuPont set (Gardiner *et al.* 2004). The random ESTs were screened by overgo-hybridization against the maize B73 BAC library (Gardiner *et al.* 2004), and only ESTs that hybridized to a single BAC contig were used for SNP discovery (Wright *et al.* 2005). SNP genotyping was performed using the Sequenom MassARRAY System (Jurinke *et al.* 2002). A complete list of the SNPs used in this study is included in Table S1 (Supporting Information). Sequence alignments, genotypes, and SNP context sequences are available at http://www.panzea.org. As in most other studies, our SNP markers were ascertained in a limited set of individuals and their absolute frequencies may therefore be skewed towards medium frequency variants. Absolute values of heterozygosity reported in our study should therefore not be compared to other published values, although relative values and differentiation are not likely to be affected by ascertainment.

Genotypes were obtained from Sequenom Inc. for 599 submitted SNPs. From this data set, quality control removed 109 SNPs: 88 SNPs had greater than 10% missing data, 15 SNPs showed extreme genotypic frequencies ($F_{IS} > 0.4$ or <-0.25), and 6 SNPs either failed, were found to be duplicates, or were typed as heterozygous in the inbred control line W22. An additional 22 SNPs were monomorphic in our sample. Nine plants were deleted from the final data set because they had greater than 20% missing data. Our final data set thus consisted of 955 individuals genotyped for 468 SNPs. Of these SNPs, 173 were scored in one of 34 candidate genes and 295 in one of 95 random genes. Candidate genes had an average annotated length of 1219 bp and contained 5.56 SNPs per gene, while random genes averaged 525 bp with 3.14 SNPs per gene.

© 2010 Blackwell Publishing Ltd

Diversity and linkage disequilibrium

We calculated Nei's gene diversity (Nei 1973) and Wright's inbreeding coefficient F_{IS} (Wright 1951) and fixation index F_{ST} (Weir 1996, p. 167), for random and candidate SNPs in both the Hill and Mound populations. The physical positions of SNPs were mapped by blasting context sequences against the maize genome (http://www.maizesequence.org), and are listed in Table S1 (Supporting Information). Using physical distances from the reference genome pseudomolecules, we then estimated the decay of linkage disequilibrium (r^2) following Remington *et al.* (2001).

Principal component analysis and genetic differentiation

Genetic differentiation within and between populations was analysed using principal component analysis (PCA) on the SNP genotype data (Patterson *et al.* 2006). This approach detects and summarizes differentiation by identifying a set of successive orthogonal principal components (PCs), each explaining a higher amount of the total variation than any remaining PCs. Significance is determined by comparing eigenvalues corresponding to successive PCs to the theoretical Tracy–Widom (TW) distribution of largest eigenvalues (Tracy & Widom 1994). As this distribution assumes independence of markers, we applied a correction for linkage among SNPs within the same gene. We used a two-stage approach in which we performed a separate PCA for each of g genes, and replaced the s_i SNPs within a gene by $s_i - k_i$ principal components, or eigenSNPs (Lin & Altman 2004), where k_i represents the number of eigenSNPS accounting for <0.1% of total variance within a gene. These $m = \sum_i^g (s_i - k_i)$ eigenSNPs were normalized by their standard deviations, and joined into a $n \times m$ matrix U, with n being the number of individuals. PCA was then performed on this matrix, with the significance of PCs determined by comparing the standardized eigenvalues of the covariance matrix of U to the TW distribution (Patterson *et al.* 2006).

Because the columns of U are normalized by their standard deviations, performing PCA on this matrix is technically equivalent to an eigenvalue decomposition of a correlation matrix. Strictly speaking, Tracy–Widom theory applies to covariance matrices, so the distribution of our eigenvalues is not guaranteed to exactly match the TW distribution (Johnstone 2001). We therefore tested the validity of our two-stage approach on 1000 data sets simulated from the observed variance and covariance between SNPs. A matrix U was calculated for each data set and the distribution of the largest standardized eigenvalues compared to the TW distribution.

Genotypes were clustered based on the significant principal components (Paschou *et al.* 2007). We applied Ward's hierarchical clustering algorithm to the matrix of Euclidean distances, calculated from the PCs, and used the R function cutree to assign individuals to each cluster.

Analysis of spatial autocorrelation and differentiation

Spatial autocorrelation of relatedness was performed using the programme SPAGeDi (Hardy & Vekemans 2002). Average pairwise genetic relatedness f_{ij} was calculated for different distance classes according to Ritland (1996) and plotted as a function of the Euclidean distance between individuals. Spatial autocorrelation of the main within-population PCs, corresponding to local differentiation, was calculated by Moran's I. Significance of spatial autocorrelation was evaluated by a Mantel test.

The spatial distribution of values of significant PCs and clustering results was visualized by means of Voronoi mosaics of the sampling coordinates, using the *tripack* package in R (R Development Core Team 2009). Spatial clustering patterns were quantified by calculating the correlation between the matrix of geographical distance and the pairwise matrix of cluster identity, in which each element takes a value of 1 for individuals from the same cluster and 0 otherwise. To evaluate spatial changes in PCs across each site, average PC values were calculated over sections spaced at 15-m intervals defined by the orthogonal intercepts of each point with a central line drawn through each field. Similarly, genetic differentiation with respect to the total population was calculated for each section as $F_{ST} = (p - p_i)^2 - (p_i(1 - p_i)/2n)/p(1 - p)$, where p is the average allele frequency in the population, p_i is the frequency within a section, and n is the number of plants sampled in that section. Significance of trends of PC values was tested by linear regression of individual plant PC values against the position along the central line. Correspondence between PC values, F_{ST}, and inferred clusters was studied by calculating the percentage of individuals belonging to each cluster for each 15 m section across the central line.

Selection analysis

It seems feasible that variation at candidate SNPs—found in genes whose mutant alleles have effects on plant phenotypes such as morphology, flowering time and seed physiology—could influence local adaptation. Accordingly, we asked whether candidate genes might

© 2010 Blackwell Publishing Ltd

be more likely to show evidence for selection than genes chosen at random and without known effects on phenotype. We tested this idea by comparing the contribution of candidate and random SNPs to population differentiation. A mixed model was fitted to the per gene median eigenSNP loadings to a single PC, with gene class as a fixed effect and number of SNPs per gene as a random term. Significance was tested by comparing the likelihood with a reduced model without the gene class term. To evaluate evidence of selection on all SNPs irrespective of gene class, we performed an outlier analysis of per-locus differences in allele frequencies between pairs of populations (Price *et al.* 2009). Outliers were identified using Quantile–Quantile (Q–Q) plots of Price *et al.*'s test statistic against a chi-square distribution with 1 degree of freedom.

Results

Diversity and linkage

Mean gene diversity was identical in both Hill and Mound sites (H_E = 0.29 ± SE 0.007). The two sites showed low but significant genetic differentiation (F_{ST} = 0.017, ±SE 0.001), similar to that reported for Balsas teosinte by Moeller *et al.* (2007). Genotype frequencies in the full data set deviated little from equilibrium expectations (F_{IS} = 0.085 ± SE 0.004), also consistent with previous reports in teosinte (Doebley *et al.* 1984); the equilibrium selfing rate estimated from F_{IS} (Haldane 1924) was ∼16%. F_{IS} was higher in the Mound site (0.088 vs. 0.065, *t* test, P = 0.001). Though we observed linkage among SNPs within a locus (≤10 kb distance, mean pairwise r^2 = 0.23) linkage disequilibrium decayed extremely rapidly (Fig. S1, Supporting Information), and SNPs in different loci were effectively unlinked (>10 kb, mean r^2 << 0.01).

Genetic differentiation

Because multiple SNPs were typed for most loci, we anticipated that analysis of population structure would be confounded by correlations among linked SNPs. Model-based methods for genetic structure analysis such as STRUCTURE (Pritchard *et al.* 2000) are not recommended for use with tightly linked markers (Falush *et al.* 2003), and our simulations showed that PCA methods consistently overestimate the number of clusters unless linkage is accounted for (Fig. S2, Supporting Information). Consistent with this, PCA on the matrix of normalized SNPs in both populations yielded 64 significant PCs, but correcting for linkage using our two-stage PCA approach (see section 'Methods') resulted in the detection of only four significant PCs in the total data set (P = 8.0e-09, 2.1e-3, 6.2e-3, 2.3e-2, 5th PC: 0.45). All further PCA results utilized the eigenSNP matrix from this two-stage approach. The first eigenvalue accounted for 46% of the significant variation (1.6% of the total variation) while the remaining three eigenvalues each contributed 18% (0.6% of the total variation). Clustering based on the first PC clearly separated the Hill and Mound sites, leaving only five individuals (1%) misclassified at each site (Fig. S3, Supporting Information). The second and third PCs further separated the Mound site into three subgroups. The fourth PC was associated with three groups within Mound as well as an apparent subdivision within the Hill plot.

To maximize the power to detect within-population structure, we performed separate principal component analysis for each site. All subsequent results are based on within-population PCs. Within the Hill site, none of the PCs were significant, suggesting little evidence for substructure. In contrast, analysis in the Mound population showed three significant PCs, indicating the presence of four differentiated groups. F_{ST} calculated over the four inferred subpopulations was 0.012, significantly higher than observed in 1000 random assignments of individuals to groups (P < 0.001, mean F_{ST} = 0.003). Mean F_{IS} within Mound subpopulations was 0.077, compared to 0.087 in the total population. Although this was still higher than the value of 0.065 observed in the Hill population, the difference was no longer significant (*t*-test, P = 0.075).

Spatial autocorrelation of relatedness

Both the Hill and Mound sites showed significant spatial autocorrelation of relatedness (Fig. 2a). Overall relatedness was low however, and few close relatives (sibs and half-sibs, f_{ij} > 0.25) were found (5 pairs in Hill, 14 pairs in Mound). Close relatives did not cluster spatially, although they were closer on average than random individuals (Hill: 31 m vs. 69 m, Mound: 76 m vs. 106 m). The Mound site showed higher relatedness over the first 50 m, but relatedness fell to 0 beyond this distance in both populations. Spatial autocorrelation calculated within the four subpopulations of the Mound site yielded a curve that more closely resembled the Hill site.

Spatial autocorrelation of principal components

Analysis of Moran's I of individual PCs (Fig. 2b–d) showed that the genetic differentiation within the Mound plot as revealed by the three significant PCs was indeed spatially structured: autocorrelation extended to the maximum distance class for the first PC, while the second and third component showed

© 2010 Blackwell Publishing Ltd

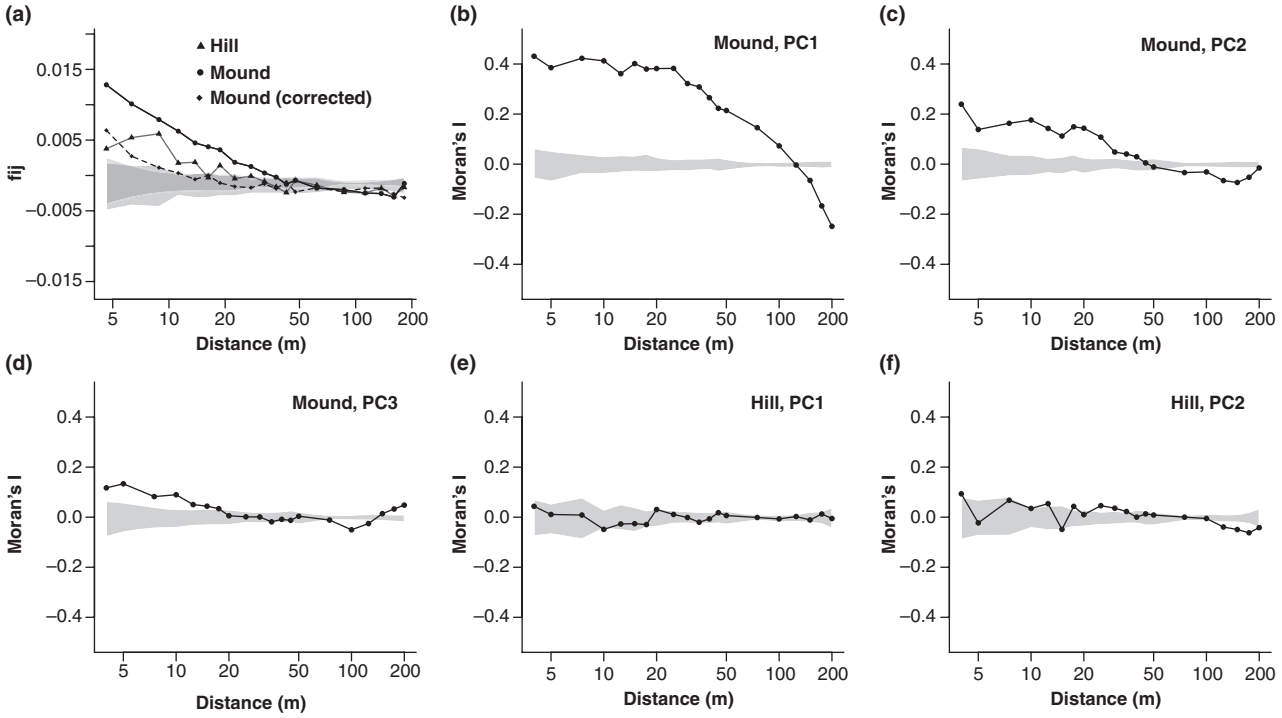

Fig. 2 Spatial autocorrelation analysis. The top left panel shows the spatial autocorrelation of relatedness for the Hill population, the Mound population and the Mound population corrected for inferred within-population structure. The remaining panels show Moran's *I* as a function of distance for the first (nonsignificant) within-Hill PC and the first three (significant) within-mound PCs. Shaded areas represent the 95% confidence envelope generated from 100 random permutations of the data.

significant autocorrelation but at decreasing spatial scales. The first two within-Hill PCs did not show significant autocorrelation, although the correlation with distance was somewhat stronger for the second PC (Fig. 2e, f).

A Voronoi heat map of the PCs confirmed these findings: clear spatial patterns were visible for the first two significant within-Mound PCs, but no spatial pattern is apparent for PCs in the Hill site (Fig. 3). The first PC in the Mound population showed an increasing trend from north to south that was accentuated at the extremes of the sampling area, while the second PC decreased from north to south, but increased again in the southern extreme of the site. No obvious pattern was evident for the third within-Mound PC.

We further confirmed spatial genetic differentiation within the Mound population by analysing the distribution of clustering results based on the three significant PCs. Allowing for three groups, three spatial clusters were apparent that roughly corresponded to the northern, central and southern parts of the field, although individuals assigned to the three clusters were found throughout the field (Fig. 4). Adding a fourth group resulted in the subdivision of the northern cluster, but without any clear spatial interpretation. Mapping clusters derived from the first three PCs in the Hill did not reveal any spatial pattern (Fig. 4), consistent with the lack of significance for the PCs.

To evaluate the sensitivity of spatial clustering to reduction in the number of markers, we compared the observed correlation between cluster identity and geographic distance in the Mound population (0.157 vs. −0.001 in Hill) to estimates from analyses performed on subsets of our marker data. A clear positive relationship is evident between number of markers analysed and the correlation between cluster identity and distance (Fig. S4, Supporting Information), with very little correlation remaining when using fewer than 250 SNPs. Also, reanalysing the mound data without correction for linkage led to a lack of correlation between cluster identity and distance (correlation coefficient −0.015).

We studied the spatial relationships between the within-plot principal components, clustering results, and levels of differentiation by plotting section-averaged PC and F_{ST} values on the relative predominance of each of three clusters (Fig. 5). The results for the Hill site again were consistent with a lack of spatial structure. The first PC showed no trend along the field mid-line, and although values of the second PC changed significantly across the field ($P < 0.001$), distance along the mid-line explained only 3% of total variation. Similarly, cluster membership based on the first three

© 2010 Blackwell Publishing Ltd

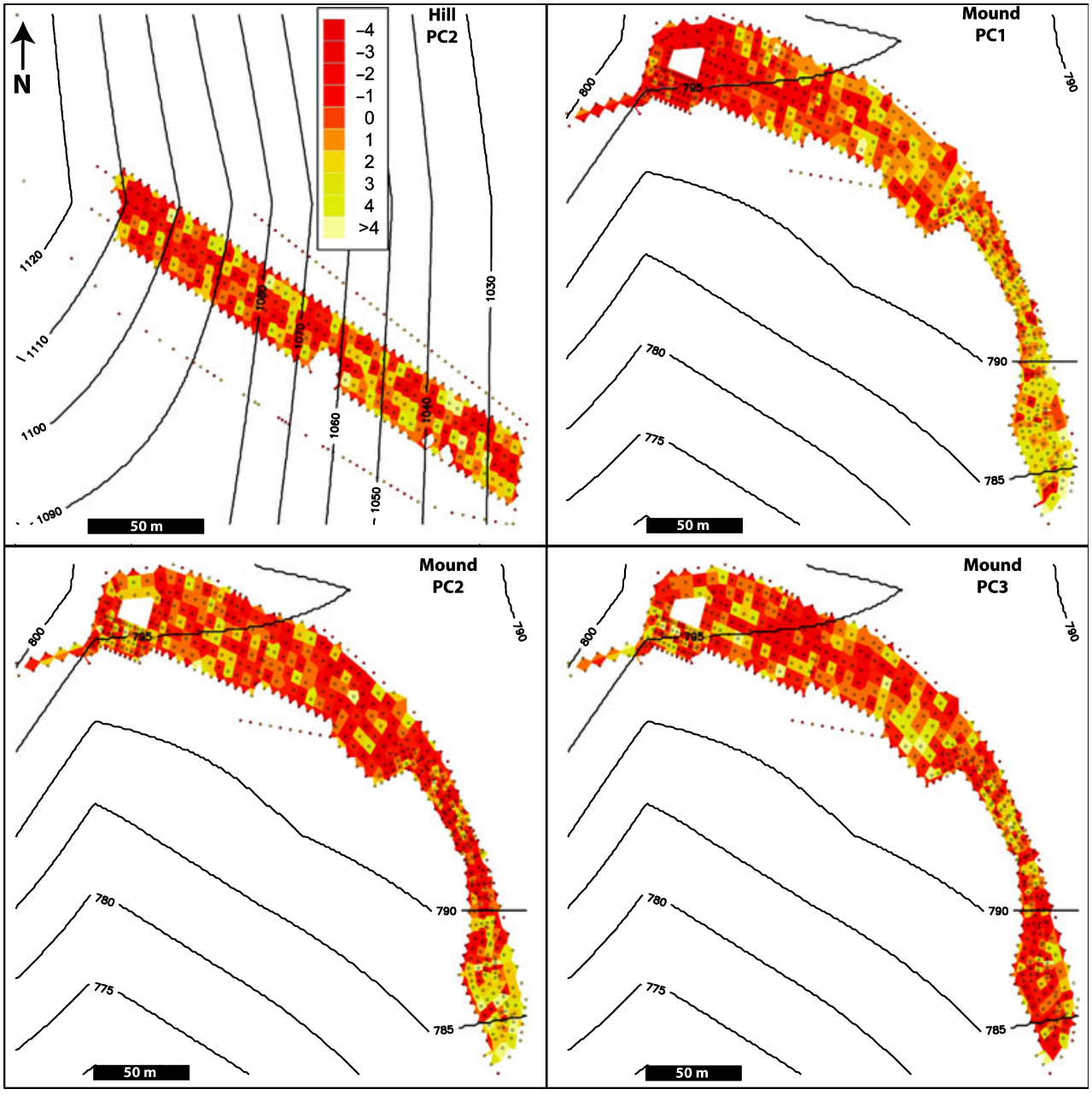

Fig. 3 Voronoi mosaics showing the distribution of PC values for Hill and Mound. White areas within each population mark patches of nonhabitat that did not contain any accessible plants.

PCs showed no spatial trend, and F_{ST} between each section and the total population was correspondingly low across the field with no clear differences between sections. Within the Mound site, however, the first PC showed a clear upward trend across the field ($P < 2e\text{-}16$, $R^2 = 0.35$), with a stronger slope visible along the first 100 m. The second PC showed a much weaker trend along the field mid-line ($P = 0.009$, $R^2 = 0.02$) but a clear upward trend was visible over the last 75 m of the field. Cluster membership changed along the field, with prevalence of the first and third cluster negatively correlated with PC1 and positively related with PC2. Prevalence of the second cluster was associated with intermediate values for both PCs, typical of the middle section of the field. Consistent with the observed spatial structure, section-wise F_{ST} in the Mound was higher compared to the Hill population, with the largest values occurring in the extremes of the field, corresponding to the sections of highest prevalence of the first and the third cluster. Projection of section-averaged elevation on the clustering results (Fig. 5) showed a possible relation between local topology and prevalence of the three

© 2010 Blackwell Publishing Ltd

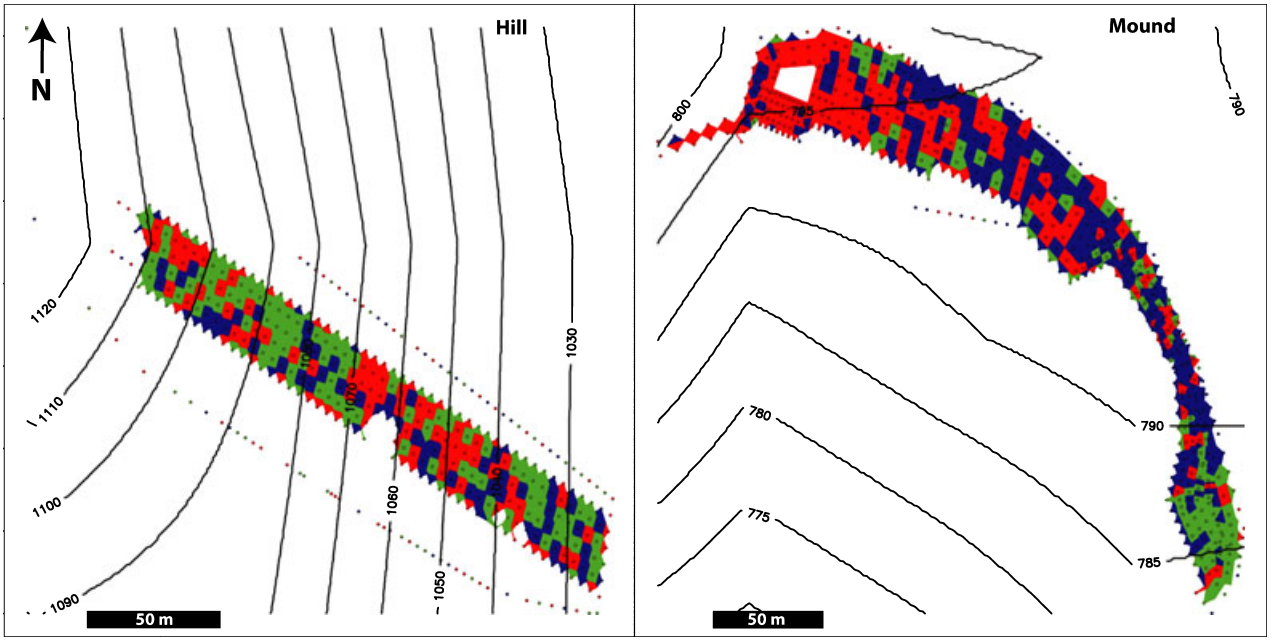

Fig. 4 Voronoi mosaics of individual assignment to each of three groups based on the first three PCs in Hill and Mound.

clusters in the Mound population: changes in slope on either side of the central, flat region corresponded to the sharpest transitions in cluster prevalence. No such relation was evident in the Hill population.

Selection

We compared the per-gene averages of eigenSNP loadings between candidate and random loci for the main significant PCs. Mixed model analysis of median loadings did not provide evidence for selection at candidate loci. Differentiation between populations and subpopulations did not differ between candidate and random SNPs. Differentiation at individual SNPs reached maximum values of F_{ST} of 0.169 (PZA03320.3) between populations and 0.152 (PZA03095.2, AY103840) within the Mound population. Q–Q analysis showed little evidence of selection at individual SNPs, however. The only deviation from the expected distribution was found within the Mound plot, where comparison between the middle- and south cluster showed 11 deviating SNPs with a maximum F_{ST} of 0.07 (Fig. S5, Supporting Information). Six of these SNPs (PZA03319.2, PZA03319.3, PZA03781.3, PZA03781.5, PZA03781.8, and PZB00232.1) mapped to three candidate loci (ath-miR156, AY107952, and BG836523), while the five remaining SNPs (PZA00407.9, PZA00471.3, PZA03094.18, PZA03095.2, and PZA03102.2) mapped to five random loci (Table S1). Removal of these 11 SNPs caused a loss of significance of the second and third within-Mound PCs but did not appreciably change the clustering results or the PC and F_{ST} trends across the field (data not shown).

Discussion

We have presented results on fine-scale spatial genetic structure in two continuous populations of a highly outcrossing species. By using dense, uniform sampling, and a large number of markers, we were able to observe extremely subtle patterns of structure within individual populations. Our study represents one of the first to provide explicit descriptions of differentiation at such a high spatial resolution. We also present one of the first applications of PCA to the study of fine-scale genetic structure and demonstrate that PCA is a powerful tool for detailed dissection of patterns of genetic diversity. Not only does it allow for the detection of significant features of genetic structure, but it provides a natural way to analyse these features spatially. Although recent studies suggest that spatial PCA results should be interpreted with caution (Novembre & Stephens 2008), our additional analyses supported the patterns found by PCA. In particular, our spatial F_{ST} results confirmed that different parts of the Mound population were indeed differentiated in terms of allele frequencies whereas no such differentiation was evident in the Hill population. We found that both the use of a large number of markers and linkage correction were necessary for detecting spatial patterns of differentiation at this scale, providing a note of caution for studies using few, or many linked markers.

We found that despite similar patterns of spatial autocorrelation of relatedness, differentiation was present in only one of the two populations. This result

© 2010 Blackwell Publishing Ltd

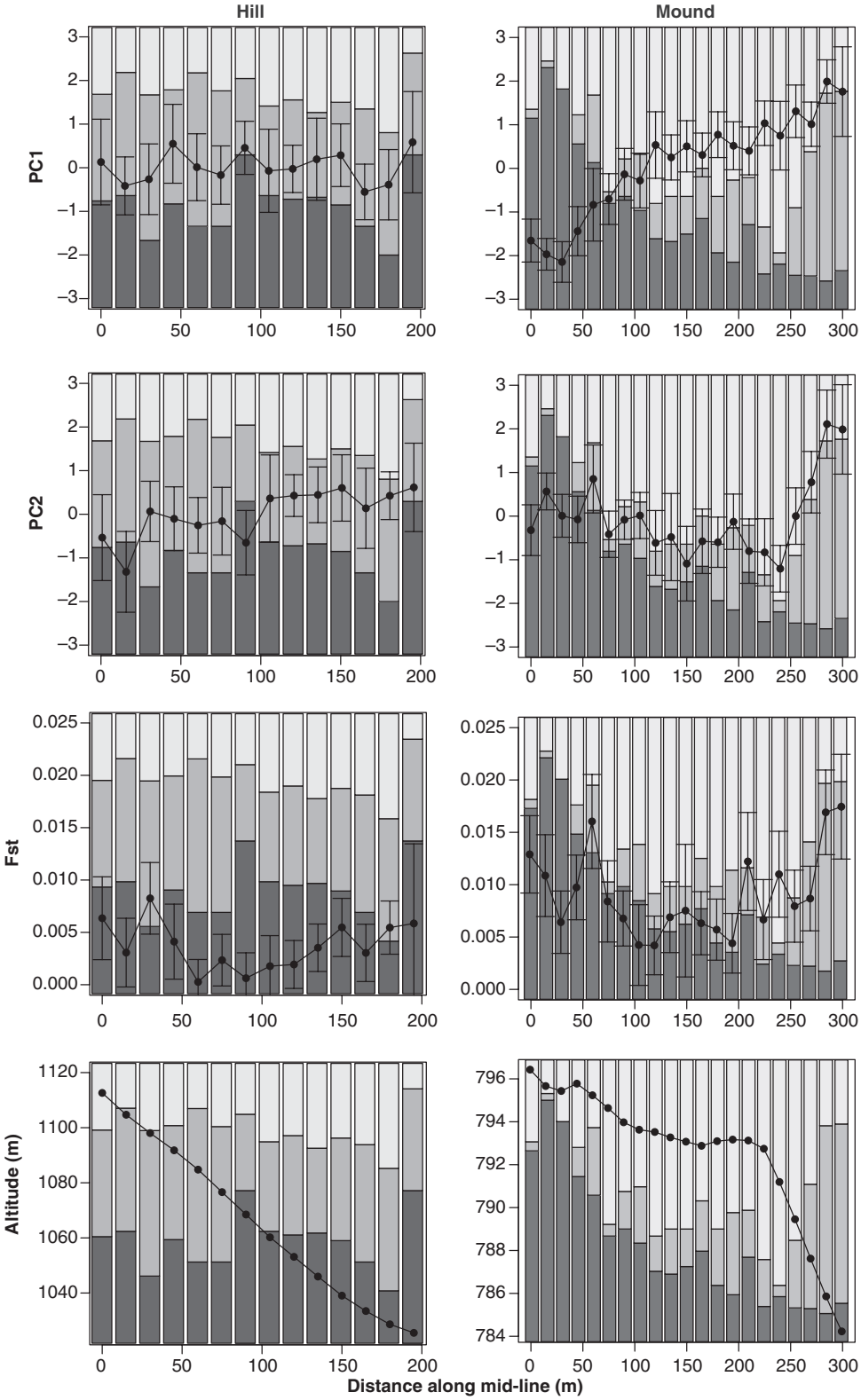

Fig. 5 Panel plot showing different parameters calculated for 15 m sections along the field midline in the Hill (left) and Mound (right) populations. From top to bottom: PC1, PC2, F_{ST} with respect to the whole population, altitude estimated with SRTM3 elevation model. Gray bars forming the background of each plot represent the proportion of individuals in each section assigned to three clusters based on the first three within-population PCs. Whiskers represent the 95% confidence intervals.

© 2010 Blackwell Publishing Ltd

suggests that subtle differences in local conditions can create contrasting patterns of genetic structure even among nearby populations of a high gene flow species. Our results on the distribution of pairwise relatedness values did not suggest a major role for kin structure in creating spatial differentiation in the Mound population. Since we genotyped single seeds, our study was most suited to detect half-sibs due to sharing of pollen donors among progeny. If pollen production were dominated by only a few plants, local differentiation might be caused by creating an excess of half-sib offspring in their vicinity. Although some potential half sib pairs were detected, their spatial distribution did not explain the spatial patterns of differentiation found in this population.

We also found little evidence suggesting a role for natural selection in explaining our data. While some studies have reported selection at a local scale under high levels of gene flow (Hamrick & Allard 1972; Antonovics 2006), neither our mixed model analysis of candidate and random SNPs, nor the distribution of per SNP values of F_{ST} provided strong evidence that selection was an important factor contributing to differentiation between and within populations of teosinte. Although some loci showed unusual levels of differentiation between subpopulations in the Mound populations, their exclusion did not remove the observed spatial patterns of differentiation.

It thus seems that nonrandom gene flow, caused by barriers to dispersal or directional environmental effects, is a more probable explanation for our results. The observed correspondence between differentiation within the Mound population and changes in slope, for example, suggested a causal role for local topography. Although the observed differences in elevation were slight, it is possible that they affect pollen and seed dispersal in the relatively flat Mound habitat, whereas the steep slope of the Hill population might preclude small elevational differences acting as barriers to gene flow. Other explanations remain, however. In the presence of prevailing wind conditions, for example, the irregular shape of the mound population could contribute to local differentiation by causing biased pollen migration into different parts of the field.

Although a definitive explanation of the patterns of differentiation observed in the Mound population will require further research, our detailed description of population structure enabled us to identify this differentiation and narrow down the possible causes. Most studies of fine scale genetic structure have exclusively studied spatial autocorrelation of relatedness, without presenting evidence for differentiation within the sampled populations. The contrasting patterns of structure observed in our two teosinte populations highlight the fact that the observation of spatial autocorrelation of relatedness may or may not be related to underlying differentiation. This result is an important one: whereas spatial autocorrelation may relate only to dispersal distances, local differentiation signals the existence of factors that restrict or direct local gene flow. It is thus difficult to interpret the biological significance of spatial autocorrelation alone, especially given its potential sensitivity to cryptic differentiation (Xu *et al.* 2006). Our paper is not the first to report on both spatial autocorrelation of relatedness and differentiation using individual genotype data. A study on *Oryza rufipogon*, a wild relative of cultivated rice (Xu *et al.* 2006), identified highly differentiated ($F_{ST} \sim 14\%$) clusters within a local population and analysed the effect of this structure on spatial autocorrelation of relatedness. The distribution of the sampled population was not continuous however, and spatial clusters coincided with discrete patches of individuals. Given that *O. rufipogon* has relatively low outcrossing rates and shows clonal reproduction, the presence of structure within the population was somewhat unsurprising. More recently, a study with similar aims to ours found trends in principal coordinates and significant spatial autocorrelation of relatedness within a continuous population of *Eucalyptus globulus* (Jones *et al.* 2007). Unfortunately, the exact nature and interpretation of differentiation in this population is difficult to gauge, as the authors did not report levels and trends of differentiation, and their analyses relied on interpolation procedures which obscure spatial patterns of individual values. The extent to which their results were affected by discontinuities in sampling or by highly localized family structure—a possibility recognized by the authors—is therefore hard to determine.

To our knowledge, our results are the first to show that spatially structured genetic differentiation may arise within continuous, densely sampled populations in the presence of high gene flow and in the absence of strong kin structure or selection. We also demonstrated for the first time how discontinuity in major PCs and F_{ST} affect the inference of significantly differentiated clusters. We showed that accounting for this differentiation can help to explain local differences in inbreeding and patterns of spatial autocorrelation of relatedness between sites that appear identical in their ecology. While the observed within-population differentiation was subtle ($F_{ST} \sim 1\%$), the fact that it is similar in scale to the differentiation seen between populations nonetheless suggests that it may be biologically relevant. Finally, in terms of marker number and sampling density, the present study represents one of the most powerful analysis of fine scale-genetic structure to date. It demonstrates how larger data sets may reveal unexpected genetic differentiation that can account for some distinct features of the data. At the same time, the

© 2010 Blackwell Publishing Ltd

relatively weak structure uncovered is probably representative of what may be expected when increasing the number of loci after failure to detect differentiation with fewer markers. Our results may therefore serve as a benchmark for future studies by showing both the possibilities and the limitations of using large data sets for describing local patterns of genetic diversity.

Acknowledgements

We would like to thank B.K. Nguyen for performing the DNA extractions and Q. Sun for determining the maize genome (AGP v1) positions of the SNPs via BLAST. We also thank O. Hardy for making the SPAGeDi source code available, and R. Cartwright for porting the code from Windows to Unix/Mac. J. Doebley would like to acknowledge funding from NSF DBI-0321467. A.J. Eckert, D. Macaya, M.R. May, L. Vann, and two anonymous reviewers provided helpful discussion on earlier versions of the manuscript.

References

Antonovics J (2006) Evolution in closely adjacent plant populations X: long-term persistence of prereproductive isolation at a mine boundary. *Heredity*, **97**, 33–37.

Barbujani G (1987) Diversity of some gene frequencies in European and Asian populations. III. Spatial correlogram analysis. *Annals of Human Genetics*, **51**, 345–353.

Beaumont MA (2005) Adaptation and speciation: what can F-st tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Bockelmann AC, Reusch TBH, Bijlsma R *et al.* (2003) Habitat differentiation vs. isolation-by-distance: the genetic population structure of *Elymus athericus* in European salt marshes. *Molecular Ecology*, **12**, 505–515.

Briggs WH, McMullen MD, Gaut BS *et al.* (2007) Linkage mapping of domestication loci in a large maize teosinte backcross resource. *Genetics*, **177**, 1915–1928.

Castric V, Bonney F, Bernatchez L (2001) Landscape structure and hierarchical genetic diversity in the brook charr, *Salvelinus fontinalis*. *Evolution*, **55**, 1016–1028.

Cavalli-Sforza L, Menozzi P, Piazza A (1993) Demic expansions and human evolution. *Science*, **259**, 639–646.

Doebley JF, Goodman MM, Stuber CW (1984) Isoenzymatic variation in *Zea* (Gramineae). *Systematic Botany*, **9**, 203–218.

Epperson BK (1995) Spatial structure of two-locus genotypes under isolation by distance. *Genetics*, **140**, 365–375.

Epperson BK, Li T (1996) Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences*, **93**, 10528–10532.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fukunaga K, Hill J, Vigouroux Y *et al.* (2005) Genetic diversity and population structure of teosinte. *Genetics*, **169**, 2241–2254.

Gardiner J, Schroeder S, Polacco ML, *et al.* (2004) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiology*, **134**, 1317–1326.

Haldane JBS (1924) A mathematical theory of natural and artificial selection. Part II: The influence of partial self-fertilisation, inbreeding, assortative mating, and selective fertilisation on the composition of Mendelian populations, and on natural selection. *Proceedings of the Cambridge Philosophical Society-Biological Sciences*, **1**, 158–163.

Hamrick JL, Allard RW (1972) Microgeographical variation in allozyme frequencies in *Avena barbata*. *Proceedings of the National Academy of Sciences*, **69**, 2100–2104.

Handley LJL, Manica A, Goudet J *et al.* (2007) Going the distance: human population genetics in a clinal world. *Trends in Genetics*, **23**, 432–439.

Hardy OJ (2003) Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Molecular Ecology*, **12**, 1577–1588.

Hardy OJ, Vekemans X (1999) Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, **83**, 145–154.

Hardy OJ, Vekemans X (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hardy OJ, Gonzalez-Martinez SC *et al.* (2004) Fine-scale genetic structure and gene dispersal in *Centaurea corymbosa* (Asteraceae). I. Pattern of pollen dispersal. *Journal of Evolutionary Biology*, **17**, 795–806.

Ingvarsson PK, Giles BE (1999) Kin-structured colonization and small-scale genetic differentiation in *Silene dioica*. *Evolution*, **53**, 605–611.

Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, **29**, 295–327.

Jones TH, Vaillancourt RE, Potts BM (2007) Detection and visualization of spatial genetic structure in continuous *Eucalyptus globulus* forest. *Molecular Ecology*, **16**, 697–707.

Jurinke C, van den Boom D, Cantor CR *et al.* (2002) The use of MassARRAY technology for high throughput genotyping. *Advances in Biochemical Engineering/Biotechnology*, **77**, 58–74.

Kalisz S, Nason JD, Hanzawa FM *et al.* (2001) Spatial population genetic structure in *Trillium grandiflorum*: the roles of dispersal, mating, history, and selection. *Evolution*, **55**, 1560–1568.

Knowles P, Perry DJ, Foster HA (1992) Spatial genetic structure in two tamarack [*Larix laricina* (Du Roi) K. Koch] populations with differing establishment histories. *Evolution*, **46**, 572–576.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Lin Z, Altman RB (2004) Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics*, **75**, 850–861.

Loiselle BA, Sork VL, Nason J *et al.* (1995) Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Malécot G (1948) *Les mathematiques de l'heredite*. Masson et Cie, Paris.

© 2010 Blackwell Publishing Ltd

Manel S, Schwartz MK, Luikart G *et al.* (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *Science*, **201**, 786–792.

Moeller DA, Tenaillon MI, Tiffin P (2007) Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics*, **176**, 1799–1809.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA*, **70**, 3321–3323.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Paschou P, Ziv E, Burchard EG *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, **3**, 1672–1686.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Piertney SB, MacColl ADC, Bacon PJ *et al.* (1998) Local genetic structure in red grouse (*Lagopus lagopus scoticus*): evidence from microsatellite DNA markers. *Molecular Ecology*, **7**, 1645–1654.

Price AL, Helgason A, Palsson S *et al.* (2009) The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genetics*, **5**, e1000505.

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

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Remington DL, Thornsberry JM, Matsuoka Y *et al.* (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences, USA*, **98**, 11479–11484.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, **67**, 175–185.

Rohlf FJ, Schnell GD (1971) An investigation of the isolation-by-distance model. *The American Naturalist*, **105**, 295–324.

Serre D, Paabo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, **14**, 1679–1685.

Slatkin M, Arter HE (1991) Spatial autocorrelation methods in population genetics. *The American Naturalist*, **138**, 499–517.

Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561–573.

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

Sokal RR, Jacquez GM, Wooten MC (1989a) Spatial autocorrelation analysis of migration and selection. *Genetics*, **121**, 845–855.

Sokal RR, Harding RM, Oden NL (1989b) Spatial patterns of human gene frequencies in Europe. *American Journal of Physical Anthropology*, **80**, 267–294.

Tracy CA, Widom H (1994) Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, **159**, 151–174.

Vigouroux Y, Glaubitz JC, Matsuoka Y *et al.* (2008) Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *American Journal of Botany*, **95**, 1240.

Weber A, Clark RM, Vaughn L *et al.* (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). *Genetics*, **177**, 2349–2359.

Weber AL, Briggs WH, Rucker J *et al.* (2008) The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. *Genetics*, **180**, 1221–1232.

Weir BS (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, MA.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Wright SI, Bi IV, Schroeder SG *et al.* (2005) The effects of artificial selection on the maize genome. *Science*, **308**, 1310–1314.

Xu X, Lu B-R, Chen YH *et al.* (2006) Inferring population history from fine-scale spatial genetic analysis in *Oryza rufipogon* (Poaceae). *Molecular Ecology*, **15**, 1535–1544.

Supporting Information

Additional supporting information may be found in the online version of the article:

Table S1 List of SNPs used in this study.

Fig. S1 Linkage disequilibrium as a function of distance in base pairs (log scale).

Fig. S2 Distribution of standardized largest eigenvalues from PCA on 1000 simulated SNP matrices with- (left) and without (right) correction for linkage. Solid lines show simulated distributions while dotted lines represent the Tracy–Widom distribution.

Fig. S3 Voronoi mosaics showing cluster assignment based on the first between population PC for the Hill (left) and Mound (right) population. Individuals assigned to the Hill population are shown in black and those assigned to the Mound population in red.

Fig. S4 Correlation between cluster identity (three within-Mound subpopulations) and geographic distance as a function of the number of SNPs used. Vertical bars represent the 95% confidence interval based on 25 random selections of loci. Results are shown for 450 to 25 loci.

Fig. S5 Q–Q plot of SNP differentiation between the middle and southern within-Mound clusters.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2010 Blackwell Publishing Ltd