Wavelet characterization of spatial pattern in allele frequency

Jesse R. Lasky[1*], Diana Gamba[1], Timothy H. Keitt[2],

1 Department of Biology, Pennsylvania State University, University Park, PA, USA
2 Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA

* lasky@psu.edu

Abstract

Characterizing spatial patterns in allele frequencies is fundamental to evolutionary biology because such patterns can inform on underlying processes. However, the spatial scales at which changing selection, gene flow, and drift act are often unknown. Many of these processes can operate inconsistently across space (causing non-stationary patterns). We present a wavelet approach to characterize spatial pattern in genotype that helps solve these problems. We show how our approach can characterize spatial patterns in ancestry at multiple spatial scales, i.e. a multi-locus wavelet genetic dissimilarity. We also develop wavelet tests of spatial differentiation in allele frequency and quantitative trait loci (QTL). With simulation we illustrate these methods under a variety of scenarios. We apply our approach to natural populations of *Arabidopsis thaliana* and traditional varieties of *Sorghum bicolor* to characterize population structure and locally-adapted loci across scales. We find, for example, that Arabidopsis flowering time QTL show significantly elevated scaled wavelet variance at $\sim 300 - 1300$ km scales. Wavelet transforms of population genetic data offer a flexible way to reveal geographic patterns and underlying processes.

Author summary

Biologists can learn about evolutionary processes by studying spatial/geographic changes in the genotype of organisms in nature. However, many previous approaches to measure spatial genetic patterns have been limited by forcing individual samples into bins of discrete size and location, hindering our ability to learn about evolution. Here we present a new continuous approach to spatial genetics that allows us to resolve patterns that change in space and opposing patterns that occur at different spatial scales.

Introduction

Since the advent of genotyping, geographic clines in allele frequencies are one of the classic patterns in evolutionary biology: common in diverse systems, driven by multiple processes, and important to understanding the maintenance of biodiversity. By characterizing patterns of spatial turnover, evolutionary biologists might infer the underlying evolutionary and ecological mechanisms. Some of the major approaches to characterizing spatial turnover include measuring the proportion of total allele frequency variation that differs between discrete populations [1, 2], calculating correlations between spatial functions and genetic variation [3, 4], and identifying geographic regions where genetic turnover is particularly high or low [5]. In recent years researchers have collected many large, spatially-distributed DNA sequence datasets in species with a wide range of life histories [6–9]. Statistical inference can be applied to these data to understand patterns of gene flow, demographic histories, and local adaptation. Although some of these approaches have a long history of use, there remain a number of thorny challenges.

For local adaptation, the selective gradients are unknown

One important force behind allele frequency clines is changing selection along environmental gradients that causes local adaptation. However, it is often not clear what environmental gradients drive local adaptation [10]. This is especially true of non-model systems and those with little existing natural history knowledge. Even for well-studied species, it is not trivial to identify the specific environmental conditions that change in space and drive local adaptation. Ecology is complex, and abiotic and biotic conditions are high-dimensional. Rather than *a priori* selection of a putative selective gradient, an alternative approach is to simply search for spatial patterns in allele frequencies that cannot be explained by neutral processes. This approach is embodied by several statistics and approaches, such as F_{ST} [11], XtX [12], spatial ancestry analysis (SPA) [4], Moran's eigenvector maps (MEMs) [3], and others.

The form and scale of spatial patterns is unknown

The functional forms (i.e. shapes) of both spatially-varying selection and neutral processes (e.g. dispersal kernels) are often unknown, as are the forms of resulting spatial patterns. For example, the specific environmental gradients driving changing selection are often not known, nor is the spatial scale at which they act, and whether they change at the same rate consistently across a landscape.

In the case of neutral processes, a homogeneous landscape approximately at equilibrium is rarely of interest to empiricists. Instead, the influence of heterogeneous landscapes and historical contingency is usually a major force behind spatial patterns in allele frequency and traits [13]. The influence of drift and range expansion can occur at a variety of spatial scales, and in different ways across a heterogenous landscape. The scale-specificity and non-stationarity of such patterns can be challenging to characterize. Estimated effective migration surfaces (EEMS) [5] are one recently developed approach to characterize non-stationary spatial genetic patterns in ancestry, identifying neighboring populations where migration appears less or greater than average.

Many approaches rely on discretization of population boundaries

Some of the aforementioned approaches rely on dividing sampled individuals into discrete spatial groups for the analysis of differences between groups. F_{ST} is one such commonly used approach, that was introduced by Wright [1] and defined as the ”correlation between random gametes, drawn from the same subpopulation, relative to the total”, where the definition of ”total” has been interpreted differently by different authors [14].

The classic approach of calculating F_{ST} to test for selection was usually applied to a small number of locations, a situation when discretization (i.e. deciding which individuals genotyped belong in which population) was a simpler problem. Current studies often sample and sequence individuals from hundreds of locations, and so the best approach for discretizing these genotyped individuals into defined 'populations' is less clear. Similarly, the EEMS approach to studying structure and gene flow still relies on discretization of samples into a populations along a single arbitrary grid [5]. However, if delineated populations are larger than the scales at which some selective gradients or barriers to gene flow act, these will be missed. Conversely, dividing samples into too small of local populations can reduce power to estimate statistics associated with each individual population. In addition to scale, at issue is precisely where to place the boundaries between populations. The problem is enhanced for broadly distributed species, connected by gene flow, that lack clear spatially distinct populations [15]. Integrating a flexible spatial scale and population boundaries into this type of analysis is the goal of this paper.

Some approaches to characterizing spatial genetic pattern are not limited by discretization, and might be generally termed ”population-agnostic” because populations are not defined. These instead use ordination of genetic loci or geographic location. Approaches that use ordination (such as PCA) of genetic loci look for particular loci with strong loadings on PCs [16] or traits with an unexpectedly high correlation with individual PCs [15]. Alternatively, ordination of distance or spatial neighborhood matrices can create spatial functions that can be used in correlation tests with genetic loci [3]. However, ordinations to create individual rotated axes are not done with respect to biology and so might not be ideal for characterizing biological patterns. For example, ordinations of genetic loci are heavily influenced by global outliers of genetic divergence [17]. The approach we present below overcomes this limitation and is not based on specific data rotations.

Wavelet characterization of spatial pattern

Instead of discretizing sampled locations into populations, a more flexible approach would be to identify localized and scale-specific spatial patterns in allele frequency. Wavelet transforms allow one to characterize the location and the scale/frequency of a signal [18]. Daubechies [18] gives a nice analogy of wavelet transforms: they are akin to written music, which indicates a signal of a particular frequency (musical notes of different pitch) at a particular location (the time at which the note is played, in the case of music). Applying this analogy to spatial genetic patterns, the frequency is the rate at which allele frequencies change in space, and the location is the part of a landscape where allele frequencies change at this rate. Applying wavelet basis functions to spatial genetic data can allow us to characterize localized patterns in allele frequency, and dilating the scale/frequency of these functions can allow us to characterize scale-specific patterns in allele frequency (see Figure S1 for an example).

Keitt [19] created a wavelet approach for characterizing spatial patterns in ecological communities. He used this approach to identify locations and scales with particular high community turnover, and applied null-hypothesis testing of these patterns. These spatial patterns in the abundance of multiple species are closely analogous to spatial patterns in allele frequency of many genetic markers across the genome, and previous spatial genetic studies have also profited by borrowing tools from spatial community ecology [20, 21]. Here we modify and build on this approach to characterize spatial pattern in allele frequency across the genome and at individual loci.

Results

Wavelet characterization of spatial pattern in allele frequency

Our implementation here begins by following the work of Keitt [19] in characterizing spatial community turnover, except that we characterize genomic patterns using allele frequencies of multiple loci in place of abundances of multiple species in ecological communities. In later sections of this paper we build off this approach and develop new tests for selection on specific loci. Wavelets allow estimation of scale-specific signals (here, allele frequency clines) centered on a given point, a, b, in two-dimensional space.

We use a variant of the Difference-of-Gaussians (DoG) wavelet function (Figure S1) [22]. The Gaussian smoothing function centered at a, b for a set of sampling points $\Omega = \{(u_1, v_1), (u_2, v_2), \ldots (u_n, v_n)\}$ takes the form

$$\eta^s_{a,b}(x, y) = \frac{k(\frac{x-a}{s}, \frac{y-b}{s})}{\sum_{(u,v)\in\Omega} k(\frac{u-a}{s}, \frac{v-b}{s})}, \tag{1}$$

where s controls the scale of analysis and $k(x, y)$ is the Gaussian kernel $k(x, y) = e^{-(x^2+y^2)/2}$.

The DoG wavelet filter then takes the form

$$\psi^s_{a,b}(x, y) = \eta^s_{a,b}(x, y) - \eta^{\beta s}_{a,b}(x, y) \tag{2}$$

where $\beta > 1$, and so the larger scale smooth function is subtracted from the smaller scale smooth to characterize the scale-specific pattern. If we use $\beta = 1.87$, then the dominant scale of analysis resulting from the DoG is s distance units [19]. This formulation of the wavelet kernel is similar in shape to the derivative-of-Gaussian kernel and has the advantage of maintaining admissibility [18] even near boundaries as each of the smoothing kernels $\eta^s_{a,b}$ are normalized over the samples such that their difference integrates to zero.

Let $f_i(u, v)$ be the allele frequency of the ith locus from a set of I biallelic markers at a location with spatial coordinates u, v. The adaptive wavelet transform of allele frequency data at locus i, centered at a, b and at scale s is then

$$(T^{wav} f_i)(a, b, s) = \frac{1}{h_{a,b}(s)} \sum_{(u,v)\in\Omega} \psi^s_{a,b}(u, v) f_i(u, v), \tag{3}$$

where the right summation is of the product of the smooth function and the allele frequencies across locations. The magnitude of this summation will be greatest when the DoG wavelet filter matches the allele frequency cline. That is, when the shape of the wavelet filter matches the allele frequency cline in space, the product of $\psi^s_{a,b}(u, v)$ and $f_i(u, v)$ will resonate (increase in amplitude) yielding greater variation in $(T^{wav} f_i)(a, b, s)$, the wavelet-transformed allele frequencies. When the spatial pattern in the wavelet filter and allele frequencies are discordant, the variation in their product, and hence the wavelet-transformed allele frequency, is reduced. Note that the sign of $f_i(u, v)$ and thus $(T^{wav} f_i)(a, b, s)$ hold no meaning to our purposes here, because we do not use information on reference versus alternate, or ancestral versus derived allelic state.

The $h_{a,b}(s)$ term in equation 3 is used to normalize the variation in the wavelet filter so that the wavelet transforms $T^{wav} f_i$ are comparable for different scales s:

$$h_{a,b}(s) = \sqrt{\sum_{(u,v)\in\Omega} [\psi^s_{a,b}(u, v)]^2} \tag{4}$$

. Below we illustrate how to apply this wavelet transform (equation 3) of spatial allele frequency patterns to characterize genome-wide patterns, as well as to test for local adaption at individual loci.

Wavelet characterization of spatial pattern in multiple loci

Researchers are often interested in characterizing spatial patterns aggregated across multiple loci across the genome to understand patterns of relatedness, population structure, and demographic history. To do so, we use

$$D^{wav}_{a,b}(s) = \sqrt{\sum_{i=1}^{I} [(T^{wav} f_i)(a,b,s)]^2} \tag{5}$$

to calculate a "wavelet genetic distance" or "wavelet genetic dissimilarity." This wavelet genetic dissimilarity is computed as the Euclidean distance (in the space of multiple loci's allele frequencies) between the genetic composition centered at a, b and other locations across s distance units. This wavelet genetic dissimilarity $D^{wav}_{a,b}(s)$ is localized in space and scale-specific. This quantity captures the level of genetic turnover at scale s centered at a, b, and is capturing similar information as the increase in average genetic distance between a genotype at a, b and other genotypes s distance units away (Figure 1E and 1F). A benefit of using the wavelet filter is that it smoothly incorporates patterns from genotypes that are not precisely s distance units away and can be centered at any location of the analysts choosing. To get the average dissimilarity across the landscape, one can also calculate the mean of $D^{wav}_{a,b}(s)$ across locations a, b at each sampled site, to get a mean wavelet genetic dissimilarity for s.

Testing the null hypothesis of no spatial pattern in allele frequency

A null hypothesis of no spatial pattern in allele frequencies can be generated by permuting the location of sampled populations among each other. Most empirical systems are not panmictic, and so this null model might be considered trivial in a sense. However, comparison with this null across scales and locations can reveal when systems shift from small-scale homogeneity (from local gene flow) to larger scale heterogeneity (from limited gene flow) [19].

Simulated neutral patterns across a continuous landscape

We conducted forward landscape genetic simulations under neutrality (or below under spatially varying selection on a quantitative trait) using the SLiM software [23], building off published approaches [24]. We simulated outcrossing, iteroparous, hermaphroditic organisms, with modest lifespans (average of ~ 4 yrs/time steps). Mating probability was determined based on a Gaussian kernel as was dispersal distance from mother [24]. Individuals became mature in the time step following their dispersal. All code is included in supplemental files. These parameters roughly approximate a short lived perennial plant with gene flow via to pollen movement and seed dispersal. Below we indicate in some figures the expected standard deviation of gene flow from the combined mechanisms, which is equal to the square root of the summed variances of each kernel (mating and propagule dispersal).

We began by characterizing a simple spatial pattern: smooth population structure and isolation by distance across continuous landscape. We simulated a square two dimensional landscape measuring 25 units on each side. In this first simulation there were only neutral SNPs. The population was allowed to evolve for 100,000 time steps before we randomly sampled 200 individuals and 1,000 SNPs with a minor allele frequency of at least 0.05. The first two principal components (PCs) of these SNPs show smooth population structure across the landscape, and that these two PCs nearly perfectly predict the spatial location of each sample (Figure S2).

We then calculated wavelet dissimilarity $D^{wav}_{a,b}(s)$ for each sampled location at a range of spatial scales s. Here and below we use a set of scales increasing by a constant log amount, which tends to result in linear increases in dissimilarity with increasing s. The mean across sampled locations for each scale was calculated and compared to the null distribution for that scale (Figure S2). The null was generated by permuting locations of sampled individuals as described above, and observed mean of dissimilarity was considered significant if it was below the 2.5 percentile or above the 97.5 percentile of dissimilarity from null permutations.

When comparing our simulated data to the null, we found that mean wavelet genetic dissimilarity was significantly less than expected under the null model at scales $s \leq 0.93$, due to local homogenization by gene flow (standard deviation = 0.28). At scales $s \geq 1.24$, wavelet dissimilarity was significantly greater than expected, due to isolation by distance, with monotonically increasing wavelet genetic dissimilarity at greater scales. These results give some intuition into how our approach characterizes spatial pattern in allele frequency.

Simulated neutral patterns in a range-expanding species

We next simulated a scenario where we expected greater heterogeneity in patterns of relatedness and genetic dissimilarity across a landscape. We simulated an invasion across a square landscape of the same size as above, but beginning with identical individuals only in the middle at the bottom edge of the landscape (Figure 1). We sampled 200 individuals at times 100, 250, 500, 1000, 1500, 2000 years, through the full populating of the landscape around 2500 years and until the 3000th year.

We characterized wavelet genetic dissimilarity across the landscape over time. There was strong heterogeneity in spatial patterns in allele frequency, demonstrated via by the variation in wavelet dissimilarity in different regions (red versus blue in Figure 1A-D). This heterogeneity in isolation-by-distance can be seen by contrasting genotypes from different regions. Near the expansion front, there is relative homogeneity and low diversity locally in new populations, but with rapid turnover in genotypes separated by space, resulting in high wavelet dissimilarity at intermediate spatial scales (Figure 1E). In the range interior, there is greater local diversity and less turnover in genotype across space, i.e. a weaker isolation by distance (Figure 1F). Supporting the role of founder effects and low diversity at expanding range margins in driving these patterns, we observed a decline in medium- and large-scale wavelet dissimilarity in later years (Fig 1G) after the landscape had been populated.

Simulated long-term neutral patterns in a heterogeneous landscape

We next simulated neutral evolution across a patchy, heterogeneous landscape, using simulated patchy landscapes (generated from earlier work) [25]. This landscape contained a substantial portion of unsuitable habitat where arriving propagules perished. We used the same population parameters as previously and simulated 100,000 years to reach approximate stability in relatedness patterns. We then calculated wavelet dissimilarity using 1,000 random SNPs of 200 sampled individuals. Wavelet dissimilarity showed localized and scale-specific patterns of low and high dissimilarity (Figure 2). Notably, the same two relatively isolated ”islands” (top left and bottom right of landscape in Figure 2) are more similar than the null at fine scales and are less similar than the null at larger scales. Stated another way, these islands have lower diversity locally (e.g. within populations) but when compared to the mainland populations they exhibit greater divergence (relative to mainland populations a similar distance apart, Figure 2). To aid interpretation of the wavelet patterns (Figure 2), we also present the first two principal components of SNPs (Figure S3), which separated each island population, respectively, from the rest of the landscape. These results highlight the capacity of the method to contrast patterns across scales in a consistent manner and only requiring dilation of the analyzing kernel, or equivalently, rescaling the spatial coordinates.

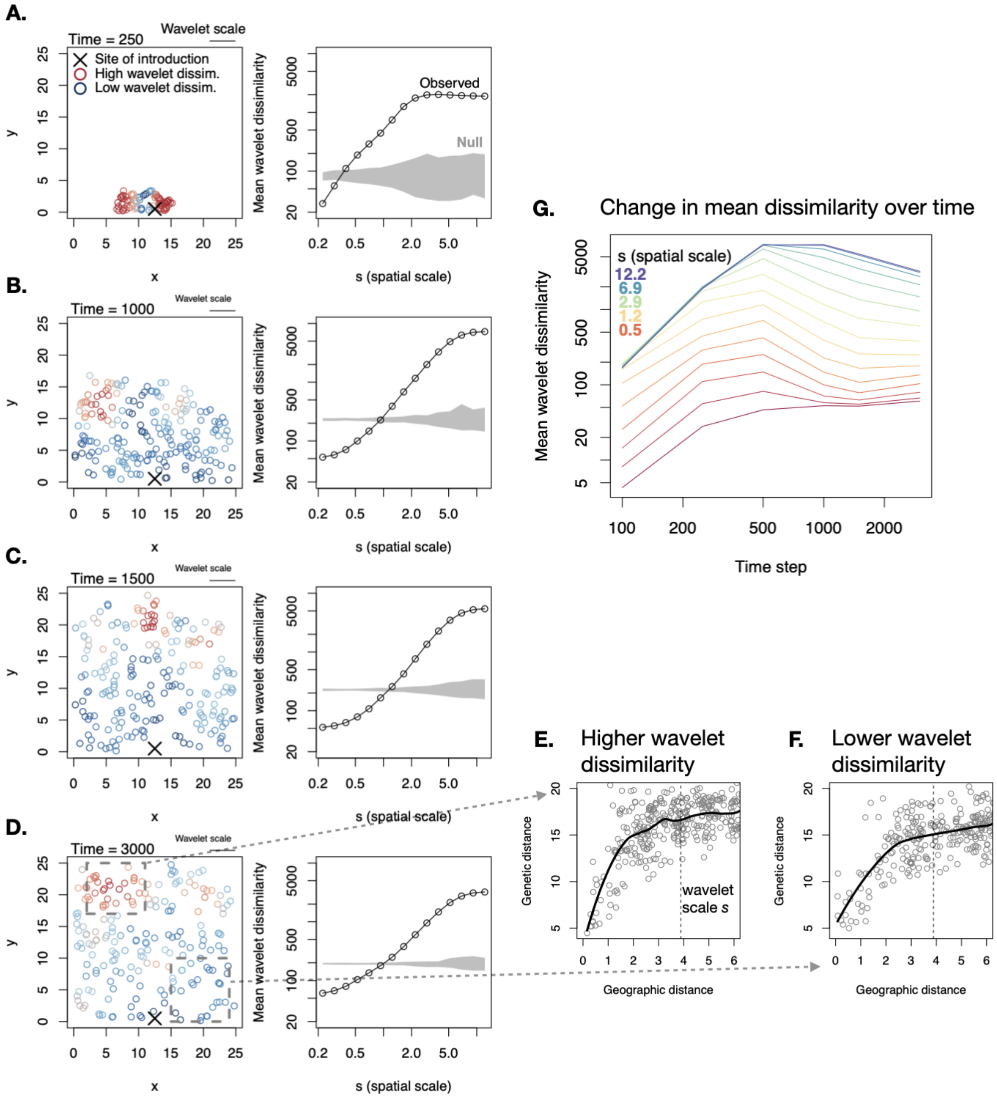


Fig 1. Wavelet genetic dissimilarity at neutral loci during an invasion across a homogeneous landscape. Wavelet genetic dissimilarity at neutral loci during an invasion across a homogeneous landscape. Left column of panels (A-D) shows a map of the landscape through time, with 200 sampled individuals at each time step and the wavelet dissimilarity at $s = 3.9$ at their location. In the last time step, 3000, two regions are highlighted (D), one with higher dissimilarity at $s = 3.9$ (E) and one with lower dissimilarity at this scale (F). (E-F) show pairwise Euclidean geographic and genetic distances for samples from these regions. These highlight the greater increase in genetic distance with geographic distance at this scale (vertical dashed lines) in (E), compared to the smaller increase in genetic distance across these distances in (F). Loess smoothing curves are shown in (E-F). (G) Mean wavelet dissimilarity across the landscape changes over time.

Finding the loci of local adaptation

Using wavelet transforms to identify outliers of spatial pattern in allele frequency

We can use our approach to identify particular genetic loci and the regions and spatial scales of turnover in allele frequency. Our strategy is to calculate $(T^{wav} f_i)(a, b, s)$ for

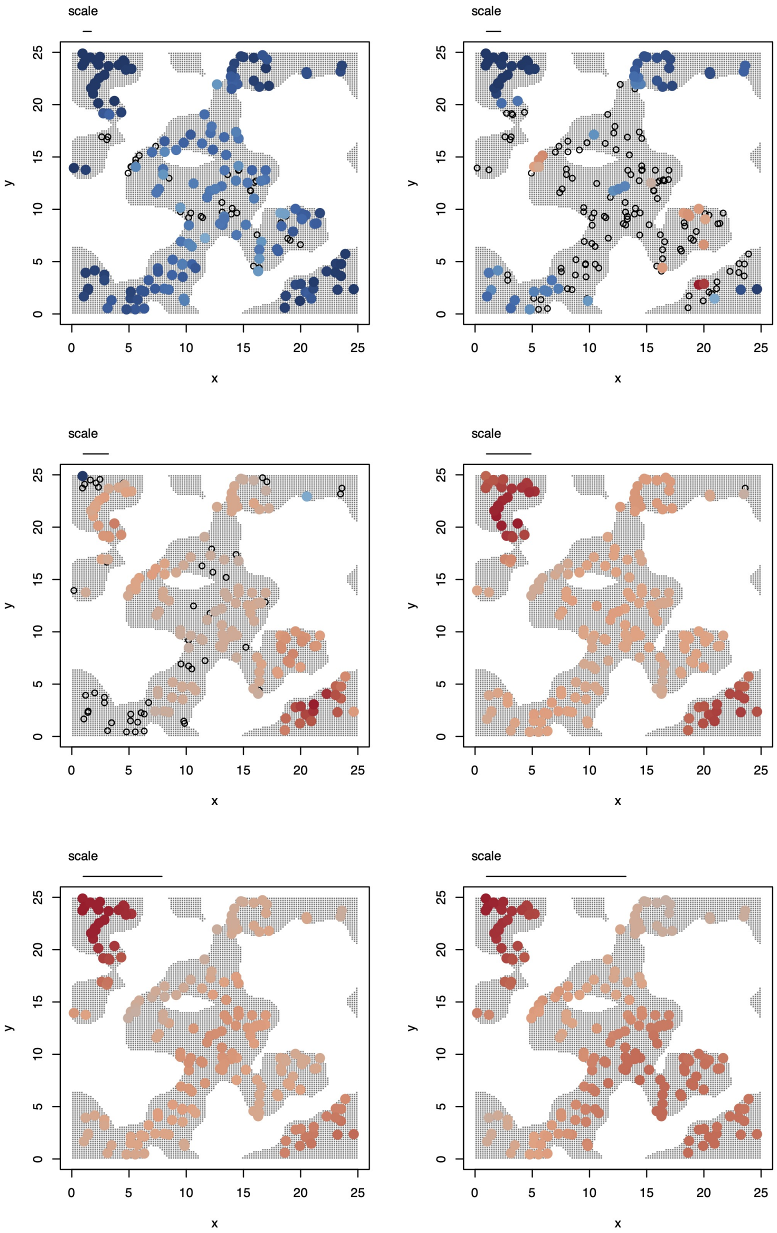


Fig 2. Neutral evolution in a heterogeneous landscape. Habitat is gray (in the background) and unsuitable areas are white. Sampled individuals are circles. Colors represent sampling locations where wavelet genetic dissimilarity was significantly high (red) or low (blue), with s, the wavelet scale, shown at top of each panel. At the smallest scales (top panels), samples are usually more similar than expected, especially in more isolated regions at lower right and upper left of the landscape. At larger spatial scales (bottom panels), all locations have significantly greater dissimilarity than expected due to limited gene flow. However, the same isolated regions at lower right and upper left of the landscape show the greatest dissimilarity at large scales (lower panels), due to their high genetic difference from the "mainland" samples at center.

each locus i at each sampling point a, b for a set of chosen spatial scales $s \in S$. Dividing the wavelet transforms of allele frequency by the standard deviation of global allele frequency variation for each locus i, $sd(f_i)$, yields a scaled measure of spatial turnover in allele frequency, $(T^{wav} f_i)(a, b, s)/sd(f_i)$, for a given location and scale. This normalization by $sd(f_i)$ results in all loci having a scaled standard deviation and variance equal to unity. This facilitates comparison of the spatial pattern of loci differing in total variance due to differences in their mean allele frequency that may arise due to different histories of mutation and drift, but arising from the same demographic processes. We then take the variance across sampling locations of $(T^{wav} f_i)(a, b, s)/sd(f_i)$, which we define the "scaled wavelet variance." This scaled wavelet variance is akin to F_{ST} in being a measure of spatial variation in allele frequency normalized to total variation (which is determined by mean allele frequency). High scaled wavelet variance for a given locus indicates high variation at that scale relative to the total variation and mean allele frequency. We then used a χ^2 null distribution across all genomic loci to calculate parametric p-values [2, 26] and used the approach of Whitlock and Lotterhos [27] to fit the degrees of freedom of this distribution to the distribution of scaled wavelet variances (see Supplemental Methods).

We simulated a species with the same life history parameters as in simulations above, with the addition of spatially varying viability selection on a quantitative trait. We imposed two geometries of spatially varying selection, one a linear gradient and the other a square patch of different habitat selecting for a different trait value. We also tested false positive rates for detecting loci under selection on the neutral patchy landscape studied above. As above with the neutral simulations, simulations began with organisms distributed across the landscape, with an ancestral trait value of zero. In these simulations, 1% of mutations influenced the quantitative trait with additive effects and with effect size normally distributed with a standard deviation of 5. For the linear gradient, the optimal trait value was 0.5 at one extreme and -0.5 at the other extreme, on a 25x25 square landscape. Selection was imposed using a Gaussian fitness function to proportionally reduce survival probability, with standard deviation σ_k. In this first simulation, $\sigma_k = 0.5$. Carrying capacity was roughly 5 individuals per square unit area [24]. Full details of simulation, including complete code, can be found in supplemental materials.

There were 3 selected loci with major allele frequency (MAF) at least 0.1 for simulations with the linear selective gradient, where the scale of mating and propagule dispersal each $\sigma = 1.1$, after 2,000 years . The two loci under stronger selection were clearly identified by $var((T^{wav} f_i)(a, b, s)/sd(f_i))$ at the larger spatial scales (Figure 3). When there is a linear selective gradient across the entire landscape, the largest spatial scale is the one most strongly differentiating environments and the strongest scaled wavelet variance was at the largest scale (Figure 3). However, power may not be greatest at these largest scales, because population structure also is greatest at these largest scales. Instead, power was greatest at intermediate scales, as seen by the lowest p-values being detected at these intermediate scales (Figure 3). At these scales there is greater gene flow but still some degree of changing selection that may maximize power to detect selection.

We next simulated discrete habitat variation, with a large central patch that selected for distinct trait values (trait optimum = 0.5) compared to the outer parts of the landscape (trait optimum = -0.5). Selection was initially weakly stabilizing ($\sigma_k = 3$ around the optimum of zero for the first 500 years to accumulate some variation, and then the patch selective differences were imposed with stronger selection, $\sigma_k = 0.08$. The scales of mating and propagule dispersal were each $\sigma = 2$. Carrying capacity was was roughly 50 individuals per square unit area.

In this simulation we present results after 3000 years, where there was a single QTL

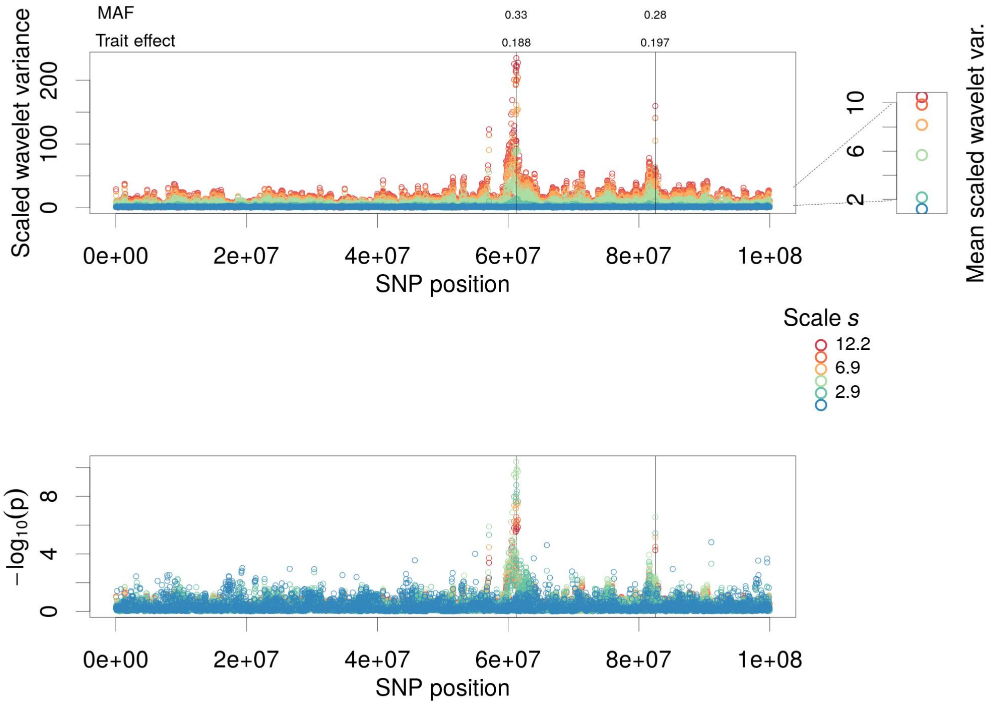


Fig 3. Scaled wavelet variance test applied to simulations with a linear selective gradient. (top panel) Genome-wide variation in scaled wavelet variance, $var((T^{wav}f_i)(a,b,s)/sd(f_i))$, for six different scales s and (bottom panel) upper-tail p-values for χ^2 test using fitted values of d.f. Each point represents a SNP at a specific scale, red is the largest scale of ~ 12.2 and blue is the smallest scale of ~ 1.6, with the intermediate scales being ~ 2.9, 5.2, 6.9, and 9.2. Simulations included a linear selective gradient and 2000 years(time steps). Loci under selection are indicated with vertical lines along with the absolute value of the derived allele's effect on the trait and MAF. At upper right the mean scaled wavelet variance across all genomic loci is shown for each scale s. The scale of mating and propagule dispersal were each $\sigma = 1.1$. Gaussian viability selection was imposed with $\sigma_k = 0.5$. Carrying capacity was roughly 5 individuals per square unit area.

under selection, with a MAF = 0.46 and the effect of one derived QTL allele on the trait = 0.497 (Figure 5). We found several spurious large scale peaks in scaled wavelet variance (Figure 4A), but when using the χ^2 test we clearly identified the single QTL under selection, with lowest p-values for intermediate scales (Figure 4B).

We then compared the application of F_{ST} to this same simulated data set (local adaptation to a single patch), using arbitrarily delineated populations. In this case, the specific delineation of populations has a major influence on whether F_{ST} can identify the selected loci. We used 'hierfstat' package in R [28] to calculate F_{ST} using the approach of Nei [29].

The most frequent locus underlying local adaptation in this simulation was only identified as a modest F_{ST} peak one of the two arbitrary grids we used to delineate populations (Figure 4D), highlighting how it can be easy to miss patterns (Figure 4C) due to these arbitrary decisions of how to subdivide samples into populations.

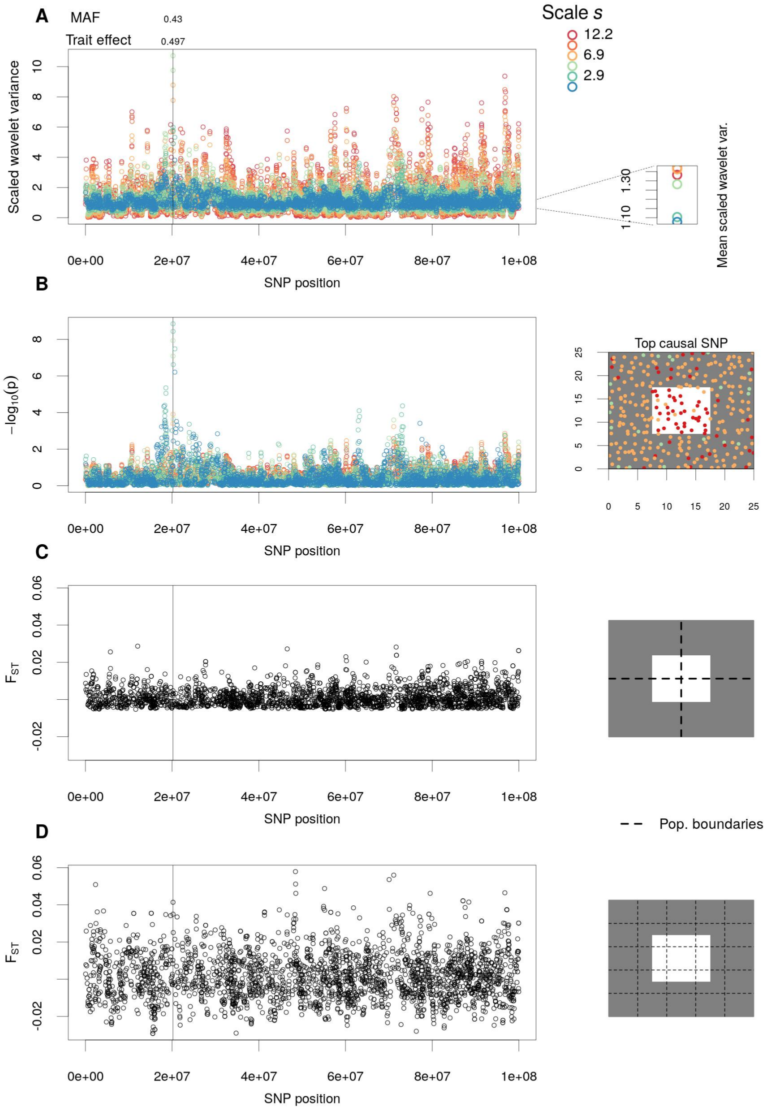


Fig 4. Scaled wavelet variance test applied to simulations with a single discrete patch of different habitat. (A) Genome-wide variation in scaled wavelet variance $var((T^{wav}f_i)(a,b,s)/sd(f_i))$ and (B) χ^2 p-values for six different scales s, for a discrete habitat difference after 3000 simulated years. Each point in the left panels represents a SNP, and wavelet statistics (A-B) at specific scales, with red being the largest scale of ~ 12.2 and blue being the smallest scale of ~ 1.6. (B) At right also shows a map of the landscape with individuals' genotypes at the causal QTL indicated with color. The locus under selection is indicated with a vertical line along with the absolute value of a derived allele's effect on the trait and MAF. (C-D) Implementation of F_{ST} using arbitrary boundaries for populations. This approach can easily miss causal loci (C) if the delineated population boundaries do not match habitat boundaries. (A) At upper right the mean scaled wavelet variance across all loci is shown for each scale s. The scale of mating and propagule dispersal were each $\sigma = 2$. Gaussian viability selection was imposed with $\sigma_k = 0.08$.

Initial evaluation of the scaled wavelet variance test

We conducted simulations on three types of landscapes with varying life history parameters as an initial assessment of the general appropriateness of the scaled wavelet variance test we proposed above. These simulations were not meant to be an exhaustive evaluation of the performance of this new test; we leave a more extensive evaluation for future studies.

Here, we again used the linear gradient landscape and the discrete habitat patch landscape but with a wider range of parameter variation. Only simulations where parameter combinations resulted in local adaptation were included. However, we also included the neutral simulations described in the previous parts of the paper to test false positive rates under these scenarios.

Overall our simulations showed good false positive rates. Across simulations and scales, the proportion of SNPs with χ^2 upper-tail $p < 0.05$ was nearly always close to but usually less than 0.05, indicating a slightly conservative test. FDR control nearly always resulted in all neutral SNPs having $q > 0.05$. Power to detect SNPs under selection ranged from low to high, depending on whether there were few SNPs or a larger number of SNPs under selection. Although selected SNPs were not all detected at $q < 0.05$, they were often closely linked to neutral SNPs that did have $q < 0.05$, though we did not consider such QTL as true positives in our conservative evaluation here. We also note that here we did not use any criteria about the distribution of selected SNPs across environments, i.e. their true role in local adaptation. Thus some of these SNPs under selection that we did not detect may have played a small role in actual local adaptation, despite their effect on the phenotype under selection (cf. [27]).

Testing for spatial pattern in quantitative trait loci (QTL)

When testing for spatially-varying selection on a quantitative trait. One approach is ask whether QTL identified from association or linkage mapping studies show greater allele frequency differences among populations than expected [30, 31]. Here we implement such an approach to compare wavelet transformed allele frequencies for QTL L to a set of randomly selected loci of the same number and distribution.

For this test we calculated the mean of scaled wavelet variance for all QTL with MAF at least 0.05 among sampled individuals (for brevity, we did not also simulate the process of mapping QTL; we leave that for future work). We then permuted the identity of causal QTL across the genome and recalculated the mean scaled wavelet variance, and repeated this process 10000 times to generate a null distribution of mean scaled wavelet variances of QTL for each scale s.

We used a similar simulation of adaptation to a square patch of habitat in the middle of a landscape. However in this case we adjusted life history parameters to result in a larger number of QTL for local adaptation (specifically we reduced the scale of the two gene flow parameters to $\sigma = 0.5$, relaxed the strength of selection so that it was now $\sigma_K = 0.5$, and reduced carrying capacity to approximately 5 individuals per square unit area).

After 1000 generations we sampled 300 individuals, from which there were 13 QTL for the trait under selection with MAF at least 0.05. We then calculated the mean scaled wavelet variance, $var((T^{wav}f_i)(a,b,s)/sd(f_i))$, for these QTL across scales s. To generate a null expectation for the mean of scaled wavelet variance for these QTL, we randomly selected 13 SNPs from the genome and recalculated mean $var((T^{wav}f_i)(a,b,s)/sd(f_i))$, and did this resampling 1000 times.

We found significantly higher mean $var((T^{wav}f_i)(a,b,s)/sd(f_i))$ for the QTL than the null expectation at all 6 scales tested. Although the scaled wavelet variance was

Landscape	K	σ	time (years)	False pos. $p < 0.05$	False pos. ($p < 0.05$, max. across scales)	False pos. $q < 0.05$	N selected SNPs	N selected SNPs $p < 0.05$	N selected SNPs $q < 0.05$
Homogeneous	5	0.2	100000	0.027	0.034	0			
Homo. colonization	5	0.2	500	0.001	0.060	0			
Homo. colonization	5	0.2	1500	0.011	0.052	0			
Homo. colonization	5	0.2	3000	0.014	0.039	0			
Neutral patchy	5	0.2	100000	0.023	0.052	< 0.001			
Selection: patch	5	0.5	1000	0.028	0.040	0	7	3	0
Selection: patch	5	0.5	2000	0.056	0.060	0	7	4	0
Selection: patch	50	2	3000	0.040	0.061	< 0.001	1	1	1
Selection: patch	25	0.1	5000	0.009	0.051	0	2	0	0
Selection: patch	25	0.1	5000	0.043	0.061	0.007	1	0	0
Selection: linear	5	1.1	2000	0.043	0.056	0	2	2	2
Selection: linear	10	2	3000	0.039	0.049	0.002	2	2	2
Selection: linear	10	2	7000	0.033	0.039	0.002	7	3	2
Selection: linear	5	0.5	3000	0.038	0.048	0	4	2	0
Selection: linear	10	0.2	3000	0.048	0.058	0	8	4	0
Selection: linear	10	1	4000	0.038	0.058	0.001	4	3	3
Selection: linear	10	2	8000	0.035	0.045	0.001	4	3	1

Table 1. Assessing the performance of our scaled wavelet variance test across a variety of simulation scenarios. Simulations are described in greater detail in the main text. K is roughly the carrying capacity per grid unit (625 total units on landscape), σ is the standard deviation of mating and propagule dispersal distance distance. We present false positive rates averaged across scales tested (or the maximum across scales where indicated). We also give the number of selected SNPs in the analyzed sample of 300 individuals with MAF at least 0.1, as well as the number of selected SNPs identified as significant in at least one scale. 6 scales were tested: ~ 1.6, 2.9, 5.2, 6.9, 9.2, 12.2, for the landscapes that were 25x25.

greatest at the largest scales for the QTL, these scales did not show as great a distinction when comparing to the null. The greatest mean wavelet variance of QTL relative to null came at the intermediate scales of 3-5, which was approximately 1/3-1/2 the width of the habitat patch (Figure 5).

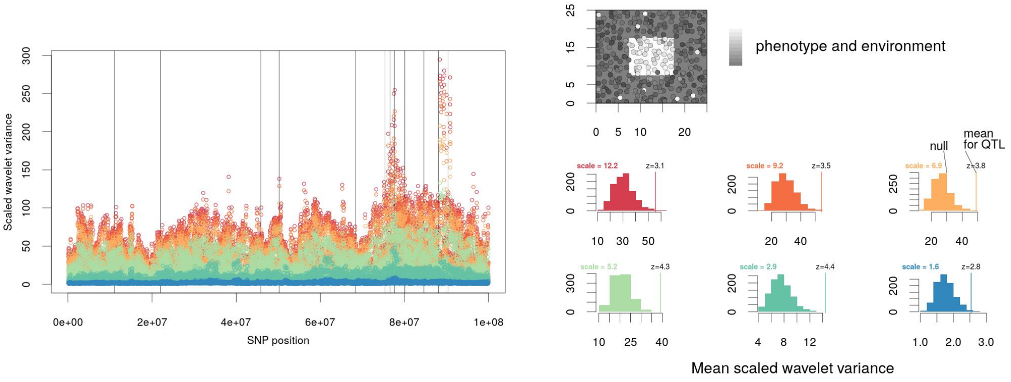


Fig 5. Testing for selection on QTL using wavelet transforms. Comparing mean scaled wavelet variance $var((T^{wav}f_i)(a,b,s)/sd(f_i))$ for QTL to that of random SNPs, for six different scales s (red = large scale and blue = small scale). Populations were locally adapted to a discrete habitat patch and results are shown after 1000 simulated time steps. QTL with MAF of at least 0.05 are indicated with vertical lines at left. The histograms at right show null distributions of mean scaled wavelet variance $var((T^{wav}f_i)(a,b,s)/sd(f_i))$ based on random samples of an equal number of markers as there were QTL (with MAF at least 0.05, n=13 here) and the observed scaled wavelet variance of QTL and its z-score.

Application to empirical systems

Genome-wide wavelet dissimilarity

We applied our approach to two empirical datasets of diverse, broadly distributed genotypes with dense marker data: 999 genotypes from 764 natural populations of the model plant, *Arabidopsis thaliana* (Brassicaceae), and 1846 traditional local varieties (landraces) from 1484 locations of the crop sorghum, *Sorghum bicolor* (Poaceae). We used a published Arabidopsis dataset [6], only including Eurasian populations, and calculated allele frequency for locations with more than one accession genotyped, resulting in 72,567 SNPs filtered for minor allele frequency (MAF$>$ 0.1) and LD. We obtained the Sorghum dataset from an integrated GBS panel of over 10K genotypes [32], from which we only included landraces, and calculated allele frequency for locations with more than one accession genotyped, resulting in 335,926 SNPs filtered for MAF$>$ 0.1. We used the R package SNPRelate [33] to generate the SNP matrix for each dataset.

For both species, we first calculated the genome-wide wavelet dissimilarity, $D^{wav}_{a,b}(s)$, across a series of increasing scales s. In both species, we observed increasing mean genome-wide wavelet dissimilarity at larger scales (Figure 6), a pattern indicative of isolation by distance, on average, across the landscape. Both species showed significantly low dissimilarity at smaller scales, likely due to the homogenizing effect of gene flow. Sorghum had significantly low dissimilarity up to the $\sim$ 47 km scale, while Arabidopsis already exhibited significantly high dissimilarity by the $\sim$ 20 km scale. This suggests the scale of gene flow due to human mediated dispersal is greater for the crop sorghum than for Arabidopsis. While both species are primarily self-pollinated, Arabidopsis lacks clear dispersal adaptations (though seeds of some genotypes form mucus in water that increases buoyancy) [34].

The specific locations of scale-specific dissimilarity revealed several interesting patterns. In Arabidopsis, at the ~ 47 km scale, there were three notable regions of significantly high dissimilarity: northeastern Iberia and extreme southern and northern Sweden (Figure 6). The high dissimilarity at this scale in northeastern Iberia corresponds to the most mountainous regions of Iberia, suggesting that limitations to gene flow across this rugged landscape have led to especially strong isolation among populations at short distances. In northern Sweden, Long et al. [35] previously found a particularly steep increase in isolation-by-distance. Alonso-Blanco et al. [6] found that genetic distance was greatest among accessions from Southern Sweden at scales from $\sim 20 - 250$ compared to some other discrete regions (though not including northern Sweden). At larger, among-region scales, dissimilarity was significantly high across the range, with Iberia and northern Sweden again being most dissimilar at ~ 619 km and joined by central Asia at ~ 1459 km as being most dissimilar. Iberia and northern Sweden contain many accessions distantly related to other accessions, likely due to isolation during glaciation and subsequent demographic histories [6]. This scale in Asia separates populations in Siberia from those further south in the Tian Shan and Himalayas, indicating substantial divergence potentially due to limited gene flow across the heterogeneous landscape.

In sorghum, there were also major differences among regions. At the ~ 263 km scale there were still regions with significantly low dissimilarity. In particular Chinese landraces had significantly low dissimilarity at this scale, potentially reflecting their more recent colonization (< 2 kya, versus e.g. ~ 5 kya for colonization of Punjab) [36, 37], and the rapid spread of closely related genotypes. By contrast, at the same ~ 263 km scale there was particularly high dissimilarity along the Rift Valley in Ethiopia, a region of high sorghum diversity and great topographic and climate heterogeneity [38], as well as in eastern India, a region where two very distinct genetic clusters meet [7]. At the among-region scale of ~ 1459 km, we found significantly high dissimilarity everywhere, especially greatest in West Africa (Burkina Faso to Nigeria), SE Africa (Zambia to Mozambique), which both correspond to the early axes of sorghum spread out of east Africa [36] and are regions of turnover in major genetic clusters [7], as well as western India and Pakistan, which corresponds to sharp rainfall gradients along which sorghum landraces may be locally adapted [38].

Identifying putative locally-adapted loci

For Arabidopsis, we focused on genotypes that were not a part of distantly related lineages (”relicts”) [6] leaving 976 genotypes, from 741 locations, for which we calculated allele frequency for locations with more than one accession genotyped. This resulted in 1,359,253 SNPs with MAF$>$ 0.1. For sorghum, we focused on landraces from sub-saharan Africa to identify putative locally adapted loci within the continent, and calculated allele frequency for locations with more than one accession genotyped, leaving 1,438 landraces from 1,094 locations and 123,334 SNPs with MAF$>$ 0.1.

The scaled wavelet variance test identified putative locally adapted loci in both species, where p-values were lower for the medium to smaller scales we tested (Figures S4 and S5). Among notable loci for Arabidopsis, the #5 SNP at the #1 locus (and 5 kb from the #1 SNP) for the ~ 282 km scale was in the DOG1 gene (Figure 7A). This SNP, Chr. 5, 18,590,327 was a peak of association with flowering time at 10$^{\text{o}}$C and germination [39] and tags known functional polymorphisms at this gene that are likely locally adaptive [39]. The spatial pattern of variation at this locus (Figure 7A) is complicated, highlighting the benefit of the flexible wavelet approach. By contrast, imposing a grid on this landscape, or using national political boundaries to calculate

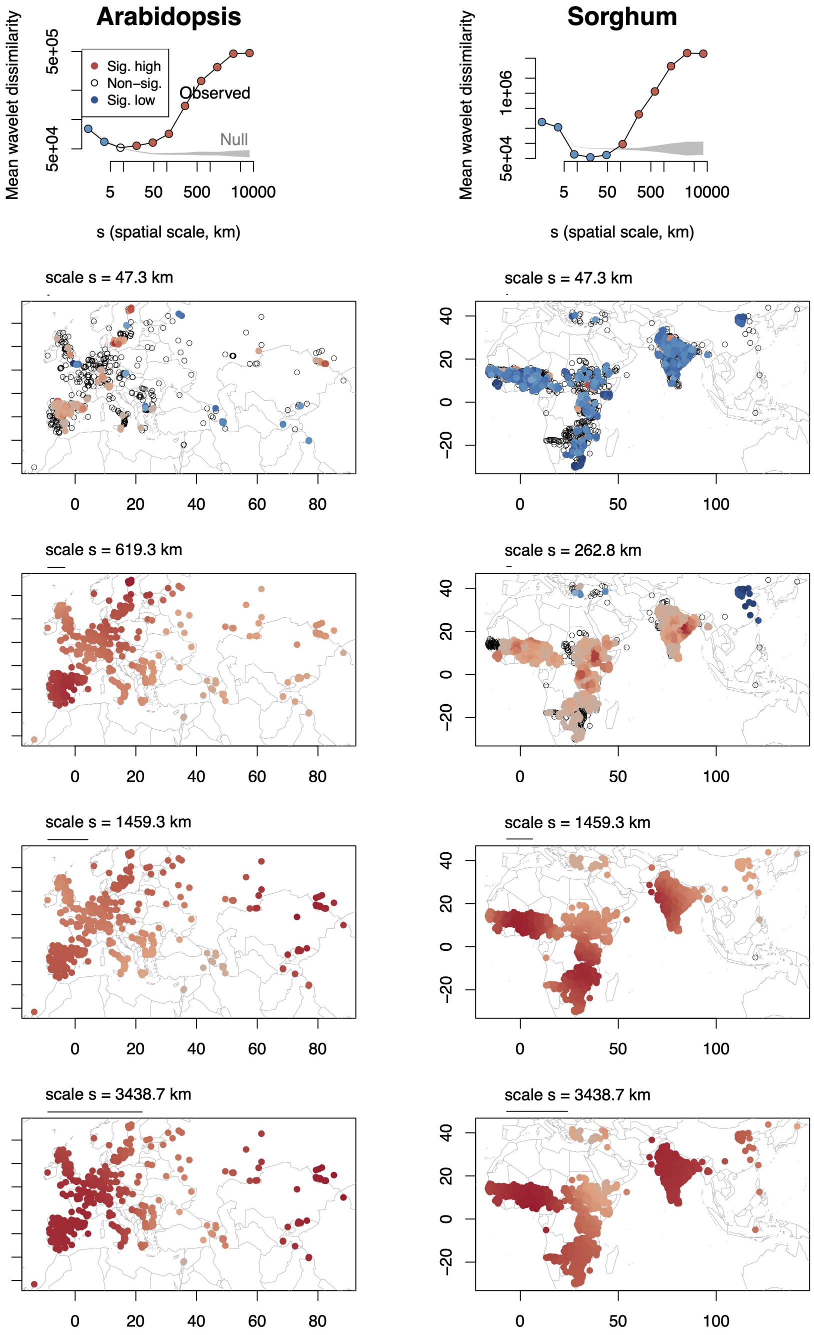


Fig 6. Genome-wide wavelet dissimilarity, $D^{wav}_{a,b}(s)$, for Arabidopsis and sorghum genotypes. Top panels show the average dissimilarity across scales compared to the null expectation. The bottom four panels show selected scales and highlight the changes is dissimilarity across locations, with each circle indicating a genotyped sample. Red indicates significantly greater wavelet dissimilarity than expected, blue significantly less than expected. For the map panels, the intensity of color shading indicates the relative variation in $D^{wav}_{a,b}(s)$ among significant locations.

F_{ST} could easily miss the signal as did Horton et al. [40]. The climate gradients driving this variation are also complicated and non-monotonic [39, 41], making it challenging for genotype-environment association approaches. At the $\sim$ 1359 km scale, the #4 locus and SNP (Figure 7B) was on chromosome 5, 648 bp upstream from SNRK2-3, which is in the family of Snf1-related kinases2 and plays an important role in signaling in response to the key abiotic stress response hormone abscisic acid (ABA) [42, 43]. This SNP was correlated to two small indels predicted to cause alternate splicing [6]: one 9 bp insertion found in 32 ecotypes (position 26711798) was always found with the SNP reference allele, and a 1 bp deletion found in 156 ecotypes was 57% of the time with the reference allele. A third insertion was present in one ecotype and overall the SNP we identified was significantly associated with the putative alternate splice variants at SNRK2-3 (Kruskal-Wallis test, $p < 10^{-5}$), suggesting we identified spatially structured functional variation in a key abiotic stress responsive signalling gene.

For Sorghum in sub-Saharan Africa, at the $\sim$ 60 km scale, the #2 locus and SNP (Chr. 6, 1,344,827 bp) was closest to ($\sim$ 18 kb distant) Sobic.006G009000, a putative calcium-activated chloride channel regulator primarily expressed in roots [44]. As expected based on the spatial scale at which this locus emerged in our genome scan, this locus showed highly heterogeneous spatial distribution, apparently much more so than expected based on the genomic distribution of SNPs (Figure 7C). Given previous evidence that sorghum landraces are adapted along relatively fine-scale soil gradients [38], we hypothesize that the pattern we detected at this locus is involved in soil adaptation. At the largest scales of $\sim 1400 - 3000$ km the #1 locus and SNP (Chr. 1, 5,016,136 bp) fell in the coding region of Sobic.001G065800, which is a glutathione S-transferase, genes that play important roles in both abiotic and biotic stressors [45]. At this locus the reference allele is nearly fixed in west Africa while the alternate allele is near fixed in southeastern Africa, regions that differ in a wide range of environmental conditions (Figure 7D).

Testing for local adaptation in quantitative trait loci (QTL)

We tested for non-random scaled wavelet variance of Arabidopsis flowering time QTL. We used previously published data on flowering time: days to flower at 10ºC measured on 1003 genotypes and days to flower at 16ºC measured on 970 resequenced genotypes [6]. We then performed mixed-model genome wide association studies (GWAS) in GEMMA (v 0.98.3) [46] with 2,048,993 M SNPs filtered for minor allele frequency (MAF> 0.05), while controlling for genome-wide similarity among ecotypes.

We found that top flowering time GWAS SNPs showed significantly elevated scaled wavelet variance at several intermediate spatial scales tested. For flowering time at both 10º and 16ºC, scaled wavelet variance was significantly elevated for the top 100 SNPs at the $\sim$ 282, 619, and 1359 km scales, but not the largest or smallest scales Fig 8. In particular the scaled wavelet variances were greatest for the $\sim$ 619 km scale, where the observed wavelet variance of QTL was 10.0 standard deviations above the mean of null permutations for 10ºC. For both temperature experiments, results were nearly equivalent if we instead used the top 1k SNPs.

Discussion

Geneticists have long studied spatial patterns in allele frequency to make inference about underlying processes of demography, gene flow, and selection. While many statistical approaches have been developed, few are flexible enough to incorporate patterns at a range of scales that are also localized in space. Because wavelet transforms have these properties, we think they may be useful tools for geneticists. Here we demonstrated

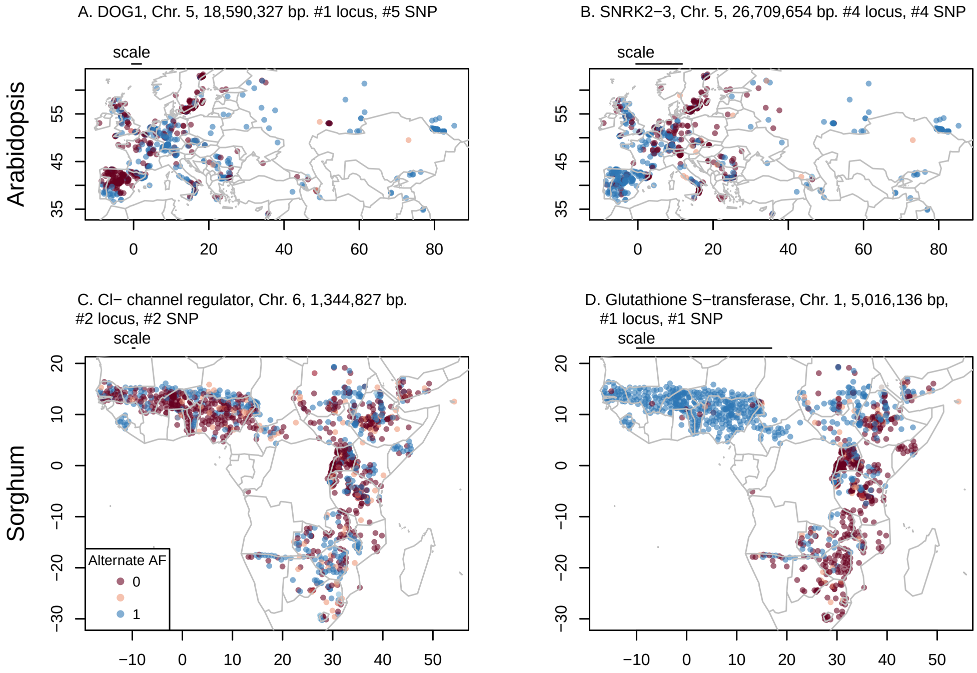

A. DOG1, Chr. 5, 18,590,327 bp. #1 locus, #5 SNP
B. SNRK2−3, Chr. 5, 26,709,654 bp. #4 locus, #4 SNP
scale
scale
Arabidopsis
55
45
35
0
20
40
60
80
C. Cl− channel regulator, Chr. 6, 1,344,827 bp. #2 locus, #2 SNP
D. Glutathione S−transferase, Chr. 1, 5,016,136 bp, #1 locus, #1 SNP
scale
scale
Sorghum
20
10
0
−10
−20
−30
−10
0
10
20
30
40
50
Alternate AF
0
1

Fig 7. SNP allelic variation (colors) that were top outliers for scaled wavelet variance test at different scales (indicated by bars above each panel). The ranks of the locus and SNP for each scale are given, where locus are defined as nearby SNPs (within 10 kb).

several applications of wavelet transforms to capture patterns in whole genome variation and at particular loci, under a range of neutral and non-neutral scenarios.

Many existing approaches are based on discretization of spatially-distributed samples into spatial bins, i.e. putative populations. However, without prior knowledge of selective gradients, patterns of gene flow, or relevant barriers, it is often unclear how to delineate these populations. For example, we can see how the specific discretization can hinder our ability to find locally-adapted loci in our simulations (Figure 4) and in empirical studies of Arabidopsis in the case of the phenology gene DOG1 that was missed in previous F_{ST} scans [6, 40].

Our goal in this paper was to provide a new perspective on spatial population genetics using the population-agnostic, and spatially smooth approach of wavelet transforms. We showed how these transforms characterize scale-specific and localized population structure across landscapes (Figures 1, 2, 6). We also showed how wavelet transforms can capture scale-specific evidence of selection on individual genetic loci (Figures 3, 4, 7) and on groups of quantitative trait loci (Figure 5 and 8). Our simulations and empirical examples showed substantial heterogeneity in the scale of patterns and localization of patterns. For example, the wavelet genetic dissimilarity allowed us to identify regions near a front of range expansion with steeper isolation by distance at particular scales due to drift (Figure 1). Additionally, we identified loci underlying local adaptation and showed an example where the evidence for this adaptation was specific to intermediate spatial scales (Figure 4). While existing approaches to characterizing population structure or local adaptation have some ability to characterize scale specific patterns, e.g. those based on ordinations of geography [3] or SNPs [15], and some can capture localized patterns (e.g [5]), there are few examples

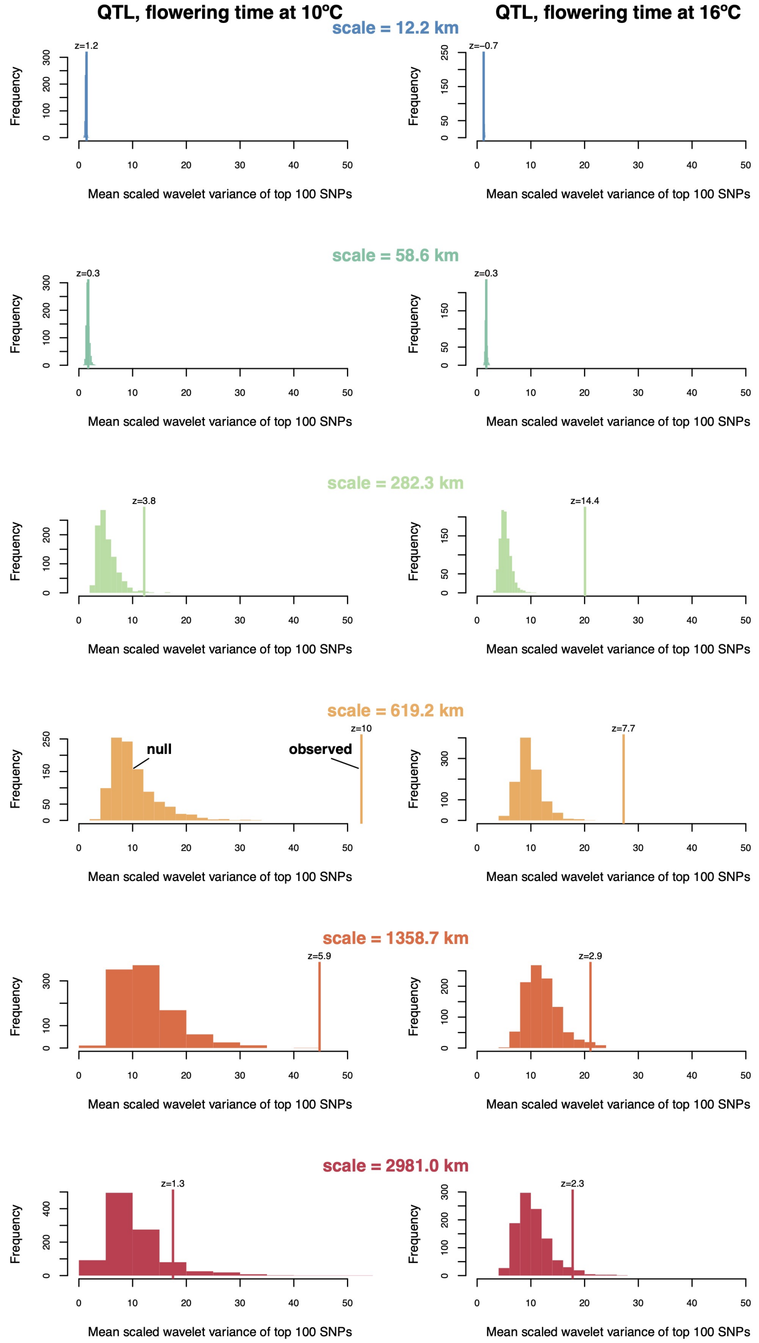


Fig 8. Testing for selection on Arabidopsis flowering time QTL. We compared scaled wavelet variance, $var((T^{wav}f_i)(a,b,s)/sd(f_i))$, of QTL with random SNPs, for five different scales s, for flowering time measured at 10ºC and 16ºC. The observed mean of the top 100 flowering time SNPs is indicated with a vertical line and a z-score. The histograms at right show null distributions of scaled wavelet variance based on permutations of an equal number of markers with an equal distribution as the flowering time QTL.

of approaches that merge both abilities. Moran's eigenvector maps (MEMs) [3] may come closest to this goal, though their scale-specificity and localization is dependent on the specific rotation of geographic axes.

The test for spatial pattern in individual loci we developed owes greatly to previous work from Lewontin and Krakauer [2] who initially developed χ^2 tests applied to the distribution of F_{ST} values, and from Whitlock and Lotterhos [27]'s approach of inferring the degrees of freedom of the χ^2 distribution using maximum likelihood and F_{ST} across loci. The χ^2 distribution underlies a number of related genetic applied across loci [47], and here shows further utility. However, we note that this test may be slightly conservative in some situations, given that we found under some demographic scenarios a failure to detect all selected loci (itself a conservative criterion for evaluating simulations) at FDR = 0.05. Nevertheless, we believe there were important signs in our work that this χ^2-based test was valuable. In particular, we found in our simulation of adaptation to a habitat patch that the scaled wavelet variance was greatest at large spatial scales but at neutral sites, which obscured spatial pattern at the causal locus (Figure 4). When applying the χ^2 test, we were able to clearly map the causal locus while spurious loci with high scaled wavelet variance fell away because spatial patterns at those loci still fit within the null distribution.

Relatedly, we found in other simulations and our empirical examples that the strongest evidence for local adaptation was often not at the largest spatial scales (Figure 8), even when the selective gradient was linear across the landscape (i.e. the largest scale, Figure 3). This enhanced power at scales sometimes smaller than the true selective gradients may be due to the limited power to resolve true adaptive clines at large scales from the genome-wide signal of isolation by distance at these scales. At intermediate scales, there may be a better balance of sufficient environmental variation to generate spatial pattern with a reduced spatial differentiation due to limited gene flow.

We note that there remain several limitations to our approach proposed here. First, the ability of wavelet transforms to capture patterns depends on the correspondence between the wavelet form (shape) and the form of the empirical patterns we seek to enhance, and there may be better functional forms to filter spatial patterns in allele frequency. Generally speaking, a more compact smoothing kernel with minimum weight in the tails will be better at revealing abrupt spatial transitions, but at the necessary cost of less precise determination of scale [48]. Smoothing kernels such as the tricube ($k_x \simeq \left[1 - x^3\right]^3$) have been shown to optimize certain trade-offs in this space and could be used to construct a difference-of-kernels wavelet. However, the overall influence of kernel shape tends to be much less than the influence of kernel bandwidth in our experience. Second, we have not yet implemented localized tests for selection (i.e. specific to certain locations) as we did with genome-wide dissimilarity. A challenge applying this test at individual loci is that there is a very large number of resulting tests from combinations of loci, locations, and scales. Therefore we have not fully exploited the localized information we derive from the wavelet transforms.

There are number of interesting future directions for research on wavelet characterization of spatial pattern in evolutionary biology. First, we could apply the wavelet transforms to genetic variation in quantitative traits measured in common gardens, to develop tests for selection on traits akin to the Q_{ST} - F_{ST} test [15, 49]. Second, we could follow the example of Al-Asadi et al. [50] and apply our measures of genetic dissimilarity to haplotypes of different size to estimate relative variation in the age of population structure. Third, we should test the performance of our tools under a wider range of demographic and selective scenarios to get a nuanced picture of their strengths and weaknesses. Fourth, null models for wavelet dissimilarity could be constructed using knowledge of gene flow processes (instead of random permutation) to identify locations and scales with specific deviations from null patterns of gene flow.

Conclusion

Population genetics (like most fields) has a long history of arbitrary discretization for the purposes of mathematical, computational, and conceptual convenience. However, the real world usually exists in shades of gray, where there are not clear boundaries between populations and where processes act simultaneously at multiple scales. We believe that wavelet transforms are one of a range of tools that can move population genetics into a richer but still useful characterization of the natural world.

Materials and methods

Simulations with SLiM

We developed our simulations by building off the spatial neutral simulation model of Battey et al. [24] and model recipes in the SLiM software [23] for spatially varying selection on quantitative traits. Parameters differed among scenarios as described previously. Additionally some parameters were consistent across all simulations. The genome had 10^8 positions, with a mutation rate of 10^{-7} and a recombination rate of 10^{-8} per generation. The fecundity of individuals in each year was a draw from a Poisson distribution with mean 0.25. Competition occurred among neighbors, potentially resulting in an increased probability of mortality above the 5% minimum probability.

Null model test for wavelet transformed patterns at individual loci

Even under neutrality, individual loci differ in their history and thus not hall have identical spatial patterns. To develop a null expectation for the distribution of scaled wavelet variance in allele frequencies across loci, we use the basic approach of Cavalli-Sforza [26] and Lewontin and Krakauer [2]. Lewontin and Krakauer [2] used χ^2 null-model tests for F_{ST} values across multiple loci. The distribution of the sum of squares of n independent standard normal variables is χ^2 with $n-1$ degrees of freedom, so that $\frac{\hat{F}_{ST}(n-1)}{\bar{F}_{ST}}$ is also χ^2 distributed with $n-1$ degrees of freedom where n is the number of populations and $\bar{F}_{ST}$ is the mean F_{ST} among loci [2]. However, the assumption of independence among variables (here, allele frequencies among populations) is often violated, and they instead are embedded in different locations in a heterogeneous (but usually unknown) metapopulation network [27, 51–53].

To solve this problem of non-independence among populations we use the same strategy that Whitlock and Lotterhos [27] applied to F_{ST}: we use the distribution of scaled wavelet variances for each locus to infer the effective number of independent populations (giving the degrees of freedom) for the χ^2 distribution. We used the [27] method: we trimmed outliers (here the bottom 2.5% SNPs for scaled wavelet variance) in scaled wavelet variance, for each scale s, then used maximum likelihood to infer the number of independent populations (using the χ^2 maximum likelihood estimation of Whitlock and Lotterhos [27]), recalculated outliers, and then refit the χ^2 distribution iteratively. Mean scaled wavelet variance was also calculated in this process while excluding SNPs in the bottom 2.5% tail as well as those with significantly high scaled wavelet variance at FDR = 0.05. We then used that estimate of the number of effective independent populations to determine the null χ^2 distribution for scaled wavelet variance.

We then used this null distribution to calculate upper tail probabilities as one-sided p-values, and then used Benjamini Hochberg FDR to get q-values. We found (like [27]) that the χ^2 distribution was sensitive to the inclusion of low MAF variants and thus we also excluded any SNPs with MAF < 0.1.

Supporting information

S1 Fig. An example of applying a difference of Gaussians (DoG) wavelet to spatial allele frequency patterns (here in one dimension). (A) shows the change in allele frequency across the spatial dimension x. (B-C) show two DoG wavelets (black curves) of two different scales s, centered at location $a = 0$, with the allele frequency pattern overlain in gray. The two selected scales (B-C) are shown because they are the scales at which variation across space in the wavelet transformed allele frequency, i.e. the product of the allele frequency and DoG, is greatest (D). These two scales capture the small scale variation in allele frequency between areas where different alleles are fixed (B), and the large scale variation between the center of the landscape where the alternate allele is present in some locations versus the edges of the landscape where the alternate allele is totally absent (C).

S2 Fig. Simulated two dimensional landscape. (A) with continuous population structure among 200 sampled individuals (circles), illustrated by the first two PCs of 1000 randomly selected SNPs (colors). In (A), each individual's color gives its SNP loadings on PC1 and PC2 according to the key at upper right. Mean of observed wavelet dissimilarities (B) among the 200 samples at a range of spatial scales s (connected by a solid black line) in comparison with the null expectation (gray ribbon) from permuted sample locations (2.5-97.5th percentiles of 100 permutations). The standard deviation of gene flow distance is indicated (dashed line).

S3 Fig. Principal component analysis on 1000 random SNPs from the neutral evolution simulation on a heterogeneous landscape. Habitat is shown as gray in the background and unsuitable areas are white. Sampled individuals are circles. Colors represent the first two PCs and show how the two populations on islands in upper left and bottom right are genetically distinct.

S4 Fig. Scaled wavelet variance test results for SNPs of Arabidopsis. Scales shown go from blue (small scale) to red (large scale), specifically, ~ 12, ~ 59, ~ 282, ~ 619, ~ 1359, ~ 2980 km.

S5 Fig. Scaled wavelet variance test results for SNPs of sorghum. Scales shown go from blue (small scale) to red (large scale), specifically, ~ 12, ~ 59, ~ 282, ~ 619, ~ 1359, ~ 2980 km.

Acknowledgments

This work was supported by NIH award 1R35GM138300-01 to JRL. We thank Emily Josephs for helpful comments.

References

1. Wright S. The genetical structure of populations;15(1):323–354. doi:10.1111/j.1469-1809.1949.tb02451.x.
2. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms;74(1):175–195.
3. Wagner HH, Chávez-Pesqueira M, Forester BR. Spatial detection of outlier loci with Moran eigenvector maps;17(6):1122–1135. doi:10.1111/1755-0998.12653.
4. Yang WY, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data;44(6):725–731. doi:10.1038/ng.2285.
5. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces;48(1):94–100. doi:10.1038/ng.3464.
6. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt K, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana;166:481–491. doi:10.1016/j.cell.2016.05.063.
7. Wang J, Hu Z, Upadhyaya HD, Morris GP. Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum;124(1):108–121. doi:10.1038/s41437-019-0249-4.
8. Machado HE, Bergland AO, Taylor R, Tilk S, Behrman E, Dyer K, et al. Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in Drosophila;10:e67577. doi:10.7554/eLife.67577.
9. Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, et al. Convergent local adaptation to climate in distantly related conifers;353(6306):1431–1433.
10. Kawecki TJ, Ebert D. Conceptual issues in local adaptation;7(12):1225–1241. doi:10.1111/j.1461-0248.2004.00684.x.
11. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure;38(6):1358–1370. doi:10.2307/2408641.
12. Gautier M. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates;201(4):1555–1579. doi:10.1534/genetics.115.181453.
13. Excoffier L, Ray N. Surfing during population expansions promotes genetic revolutions and structuration;23(7):347–351. doi:10.1016/j.tree.2008.04.004.
14. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants;23(9):1514–1521. doi:10.1101/gr.154831.113.
15. Josephs EB, Berg JJ, Ross-Ibarra J, Coop G. Detecting Adaptive Differentiation in Structured Populations with Genomic Data and Common Gardens;211(3):989–1004. doi:10.1534/genetics.118.301786.
16. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data;33(4):1082–1093. doi:10.1093/molbev/msv334.

17. Peter BM, Petkova D, Novembre J. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography;37(4):943–951. doi:10.1093/molbev/msz280.

18. Daubechies I. Ten lectures on wavelets. SIAM;.

19. Keitt TH. On the quantification of local variation in biodiversity scaling using wavelets; p. 168–80.

20. Lasky JR, Des Marais DL, McKay JK, Richards JH, Juenger TE, Keitt TH. Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate;21(22):5512–5529. doi:10.1111/j.1365-294X.2012.05709.x.

21. Fitzpatrick MC, Keller SR. Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation;18(1):1–16. doi:10.1111/ele.12376.

22. Muraki S. Multiscale volume representation by a DoG wavelet;1(2):109–116. doi:10.1109/2945.468408.

23. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model;36(3):632–637. doi:10.1093/molbev/msy228.

24. Battey C, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks;9:e54507. doi:10.7554/eLife.54507.

25. Lasky JR, Keitt TH. Reserve Size and Fragmentation Alter Community Assembly, Diversity, and Dynamics.;182(5):E142–E160. doi:10.1086/673205.

26. Cavalli-Sforza LL. Population structure and human evolution;164(995):362–379. doi:10.1098/rspb.1966.0038.

27. Whitlock MC, Lotterhos KE. Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST;186:S24–S36. doi:10.1086/682949.

28. Goudet J. hierfstat, a package for r to compute and test hierarchical F-statistics;5(1):184–186. doi:10.1111/j.1471-8286.2004.00828.x.

29. Nei M. Molecular Evolutionary Genetics. Columbia University Press;. Available from: `http://www-degruyter-com/document/doi/10.7312/nei-92038/html`.

30. Berg JJ, Coop G. A Population Genetic Signal of Polygenic Adaptation;10(8):e1004412. doi:10.1371/journal.pgen.1004412.

31. Price N, Moyers BT, Lopez L, Lasky JR, Monroe JG, Mullen JL, et al. Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana;115(19):5028–5033. doi:10.1073/pnas.1719998115.

32. Hu Z, Olatoye MO, Marla S, Morris GP. An Integrated Genotyping-by-Sequencing Polymorphism Map for Over 10,000 Sorghum Genotypes;12(1):180044. doi:10.3835/plantgenome2018.06.0044.

33. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data;28(24):3326–3328. doi:10.1093/bioinformatics/bts606.

34. Saez-Aguayo S, Rondeau-Mouro C, Macquet A, Kronholm I, Ralet MC, Berger A, et al. Local Evolution of Seed Flotation in Arabidopsis;10(3):e1004221. doi:10.1371/journal.pgen.1004221.

35. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden;45(8):884–890. doi:10.1038/ng.2678.

36. Kimber CT. Origins of domesticated sorghum and its early diffusion to India and China; p. 3–98.

37. Qingshan L, Dahlberg JA. Chinese Sorghum Genetic Resources;55(3):401–425.

38. Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, et al. Genome-environment associations in sorghum landraces predict adaptive traits;1(6):e1400218. doi:10.1126/sciadv.1400218.

39. Martínez-Berdeja A, Stitzer MC, Taylor MA, Okada M, Ezcurra E, Runcie DE, et al. Functional variants of DOG1 control seed chilling responses and variation in seasonal life-history strategies in Arabidopsis thaliana;117(5):2526–2534. doi:10.1073/pnas.1912451117.

40. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel;44:212–216. doi:10.1038/ng.1042.

41. Gamba D, Lorts C, Haile A, Sahay S, Lopez L, Xia T, et al.. The genomics and physiology of abiotic stressors associated with global elevation gradients in Arabidopsis thaliana;. Available from: https://www.biorxiv.org/content/10.1101/2022.03.22.485410v1.

42. Wang K, He J, Zhao Y, Wu T, Zhou X, Ding Y, et al. EAR1 Negatively Regulates ABA Signaling by Enhancing 2C Protein Phosphatase Activity;30(4):815–834. doi:10.1105/tpc.17.00875.

43. Cai G, Wang Y, Tu G, Chen P, Luan S, Lan W. Type A2 BTB Members Decrease the ABA Response during Seed Germination by Affecting the Stability of SnRK2.3 in Arabidopsis;21(9):3153. doi:10.3390/ijms21093153.

44. McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, et al. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization;93(2):338–354. doi:10.1111/tpj.13781.

45. Gullner G, Komives T, Király L, Schröder P. Glutathione S-Transferase Enzymes in Plant-Pathogen Interactions;9.

46. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies;44(7):821–824. doi:10.1038/ng.2310.

47. François O, Martins H, Caye K, Schoville SD. Controlling false discoveries in genome scans for selection;25(2):454–469. doi:10.1111/mec.13513.

48. Heisenberg W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik;43(3):172–198. doi:10.1007/BF01397280.

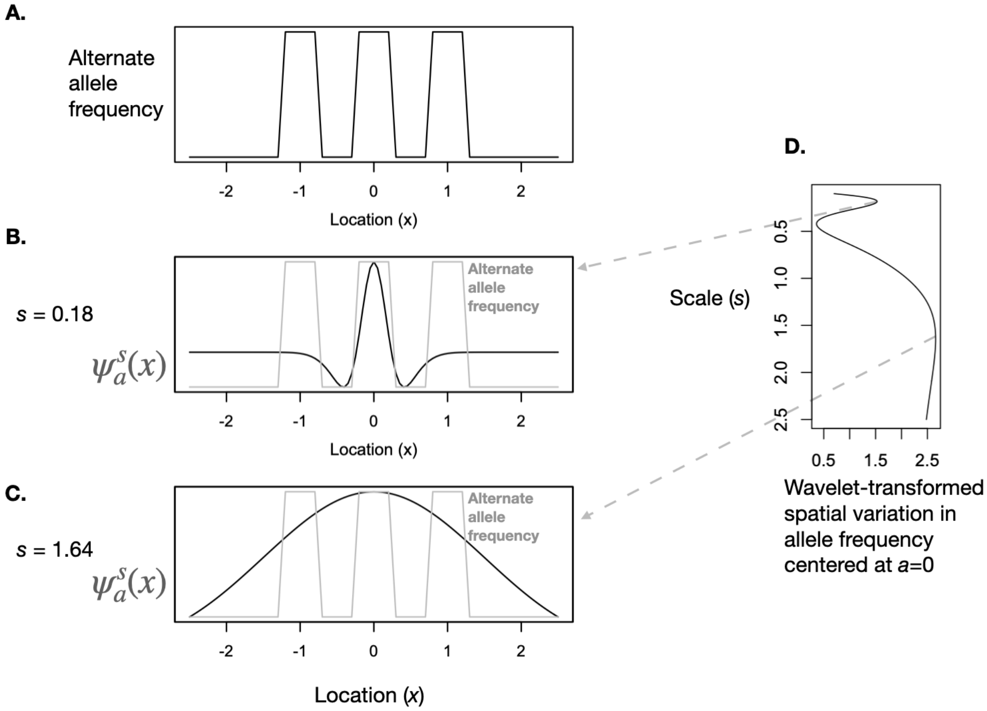
49. Whitlock MC, Guillaume F. Testing for Spatially Divergent Selection: Comparing QST to FST;183(3):1055–1063. doi:10.1534/genetics.108.099812.

50. Al-Asadi H, Petkova D, Stephens M, Novembre J. Estimating recent migration and population-size surfaces;15(1):e1007908. doi:10.1371/journal.pgen.1007908.

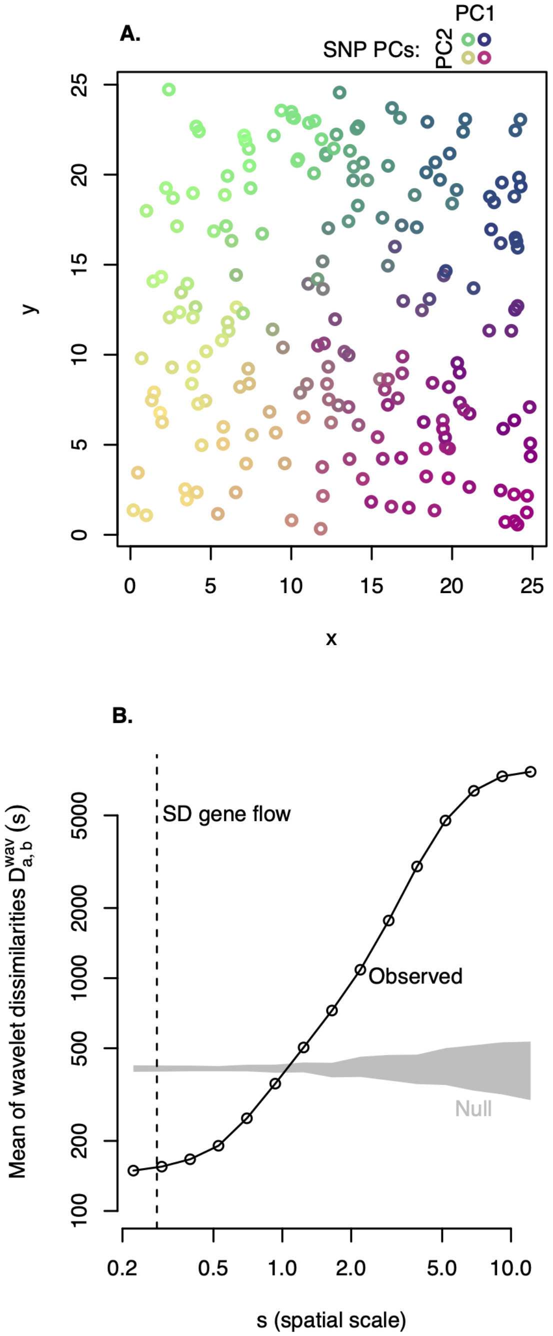
51. Nei M, Maruyama T. Lewontin-Krakauertest for neutral genes;80(2):395.

52. Robertson A. Remarks on the Lewontin-Krakauer test;80(2):396. doi:10.1093/genetics/80.2.396.

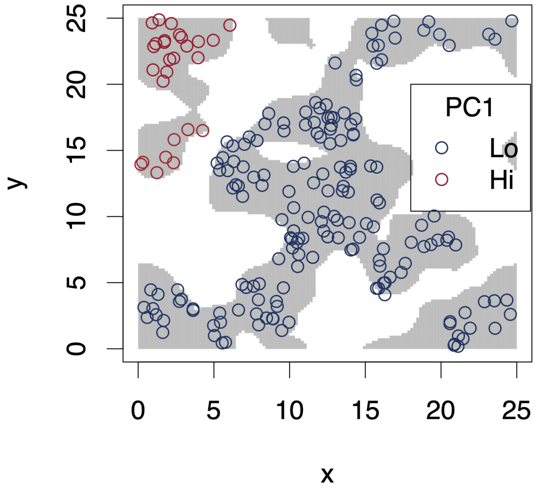
53. Lewontin RC, Krakauer J. Testing the Heterogeneity of F Values;80(2):397–398.

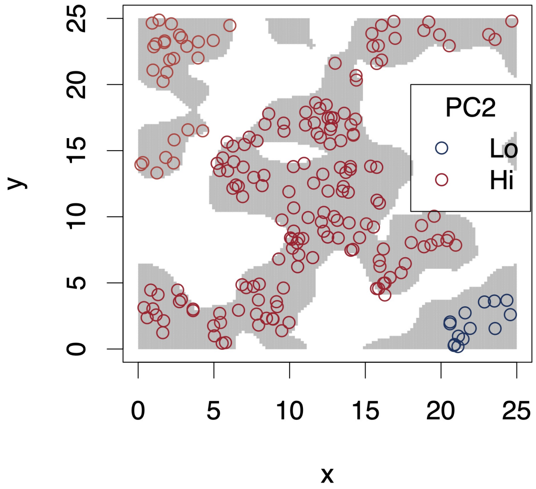


S1 Fig. An example of applying a difference of Gaussians (DoG) wavelet to spatial allele frequency patterns (here in one dimension). (A) shows the change in allele frequency across the spatial dimension x. (B-C) show two DoG wavelets (black curves) of two different scales s, centered at location $a = 0$, with the allele frequency pattern overlain in gray. The two selected scales (B-C) are shown because they are the scales at which variation across space in the wavelet transformed allele frequency, i.e. the product of the allele frequency and DoG, is greatest (D). These two scales capture the small scale variation in allele frequency between areas where different alleles are fixed (B), and the large scale variation between the center of the landscape where the alternate allele is present in some locations versus the edges of the landscape where the alternate allele is totally absent (C).

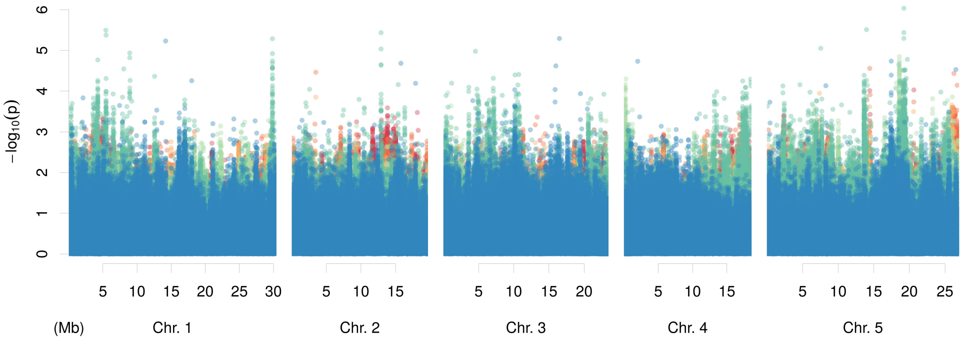


S2 Fig. Simulated two dimensional landscape. (A) with continuous population structure among 200 sampled individuals (circles), illustrated by the first two PCs of 1000 randomly selected SNPs (colors). In (A), each individual's color gives its SNP loadings on PC1 and PC2 according to the key at upper right. Mean of observed wavelet dissimilarities (B) among the 200 samples at a range of spatial scales s (connected by a solid black line) in comparison with the null expectation (gray ribbon) from permuted sample locations (2.5-97.5th percentiles of 100 permutations). The standard deviation of gene flow distance is indicated (dashed line).

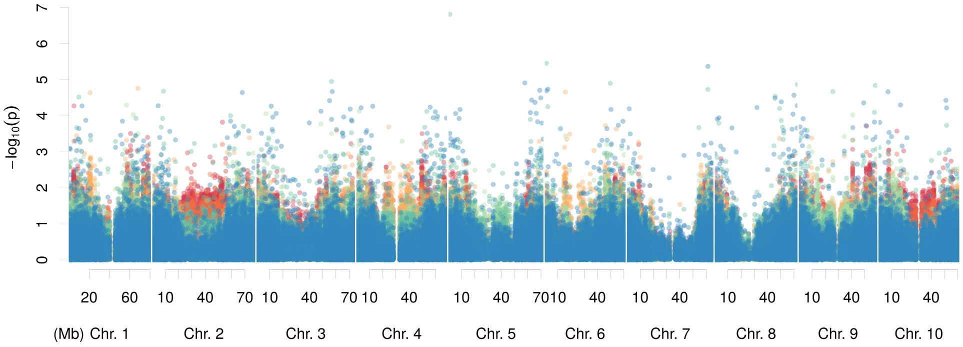




S3 Fig. Principal component analysis on 1000 random SNPs from the neutral evolution simulation on a heterogeneous landscape. Habitat is shown as gray in the background and unsuitable areas are white. Sampled individuals are circles. Colors represent the first two PCs and show how the two populations on islands in upper left and bottom right are genetically distinct.



S4 Fig. Scaled wavelet variance test results for SNPs of Arabidopsis. Scales shown go from blue (small scale) to red (large scale), specifically, ∼ 12, ∼ 59, ∼ 282, ∼ 619, ∼ 1359, ∼ 2980 km.



S5 Fig. Scaled wavelet variance test results for SNPs of sorghum. Scales shown go from blue (small scale) to red (large scale), specifically, ∼ 12, ∼ 59, ∼ 282, ∼ 619, ∼ 1359, ∼ 2980 km.