Wavelet characterization of spatial pattern in allele frequency

Jesse R. Lasky[1], Margarita Takou[1], Diana Gamba[1], and Timothy H. Keitt[2]

[1]Pennsylvania State University, Department of Biology
[2]University of Texas at Austin, Department of Integrative Biology

September 7, 2023

Abstract

Characterizing spatial patterns in allele frequencies is fundamental to evolutionary biology because such patterns can reveal underlying processes. However, the spatial scales at which gene flow, changing selection, 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 allele frequency that helps solve these problems. We show how our approach can characterize spatial patterns in relatedness 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 different scenarios. We also apply our approach to natural populations of *Arabidopsis thaliana* to characterize population structure and locally-adapted loci across scales. We find, for example, that Arabidopsis flowering time QTL show significantly elevated scale-specific genetic variance at 300 to 1300 km scales. Wavelet transforms of allele frequencies offer a flexible way to reveal geographic patterns and underlying processes.

1 Introduction

Since the advent of genotyping, geographic clines in allele frequency are one of the classic patterns in evolutionary biology: frequently observed, driven by multiple processes, and important to understanding the maintenance of biodiversity. Researchers often estimate and model spatial allele frequency patterns to make inference on underlying evolutionary and ecological mechanisms. Some of the major approaches to characterizing spatial patterns include measuring the proportion of total allele frequency variation that differs between discrete populations (Lewontin and Krakauer 1973; Wright 1949), calculating

correlations between spatial functions and genetic variation (Wagner, Chávez-Pesqueira, and Forester 2017; Yang et al. 2012), and identifying geographic regions where genetic turnover is particularly high or low (Petkova, Novembre, and Stephens 2016). In recent years researchers have collected many large, spatially-distributed DNA sequence datasets in diverse species (Alonso-Blanco et al. 2016; Machado et al. 2021; J. Wang et al. 2020; Yeaman et al. 2016). Statistical inference can be applied to these data to understand gene flow, demographic histories, and spatially-varying selection. Although some statistical approaches have a long history of use, there remain a number of thorny challenges.

1.1 The form and scale of relevant spatial patterns is unknown

Humans can infer seemingly meaningful spatial patterns in even randomly generated images (Rorschach 1921). So what are the spatial patterns we are looking for? 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 (Manel et al. 2003) and historical contingency is usually a major force behind spatial patterns in allele frequency and traits (Excoffier and Ray 2008). As a result, researchers often attempt to characterize spatial patterns of relatedness and genetic similarity to make inferences about variation in gene flow (McRae et al. 2008; Peterman 2018; I. J. Wang, Savage, and Bradley Shaffer 2009), and recent population expansion (Slatkin 1993). The influence of gene flow, 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) (Petkova, Novembre, and Stephens 2016) are one recently developed approach to characterize spatially non-stationary population structure, identifying neighboring populations where migration appears less or greater than average.

1.2 The spatially-varying selective gradients causing local adaptation 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 (Kawecki and Ebert 2004). 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 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} (Weir and Cockerham 1984), XtX (Gautier 2015), spatial ancestry analysis (SPA) (Yang et al. 2012), Moran's eigenvector maps (MEMs) (Wagner, Chávez-Pesqueira, and Forester 2017), and others.

1.3 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 (Wright 1949) 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 (Bhatia et al. 2013).

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 (Petkova, Novembre, and Stephens 2016). 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 (Emily B. Josephs et al. 2019). 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 (Duforet-Frebourg et al. 2016) or traits with an unexpectedly high correlation with individual PCs (Emily B. Josephs et al. 2019). Alternatively, ordination of distance or spatial neighborhood matrices can create spatial functions that can be used in correlation tests with genetic loci (Wagner, Chávez-Pesqueira, and Forester 2017). However, ordinations to create individual rotated axes are not done with re-

spect 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 (Peter, Petkova, and Novembre 2020) and uneven sampling (McVean 2009). The approach we present below is not based on specific rotations of axes.

1.4 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 or frequency of a signal (Daubechies 1992). Daubechies (1992) 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 could allow us to characterize localized patterns in allele frequency, and dilating the scale of these functions could allow us to characterize scale-specific patterns in allele frequency (see Figure S1 for an example). Note that wavelets are distinct from Fourier analysis. Wavelets capture localized signals because the basis functions' variance goes to zero moving away from the focal location, while Fourier can only capture global average patterns as it uses stationary (unchanging) basis functions.

Keitt (2007) 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 created 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 (Fitzpatrick and Keller 2015; Jesse R. Lasky, Des Marais, et al. 2012). Here we modify and build on this approach to characterize spatial pattern in allele frequency across the genome and at individual loci.

2 Results

2.1 Wavelet characterization of spatial pattern in allele frequency

Our implementation here begins by following the work of Keitt (2007) 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. Our implementation of wavelets allows estimation of scale-specific signals (here, allele frequency clines) centered on a given point, a, b, in two-dimensional space. We use a version of the Difference-of-Gaussians (DoG) wavelet function (Figure S1) (Muraki 1995). We start with a 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 function 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 (Keitt 2007). This formulation of the wavelet kernel is similar in shape to the derivative-of-Gaussian kernel and has the advantage of maintaining admissibility (Daubechies 1992) 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 major 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 among locations 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. For consistency, here we choose major allele frequency for $f_i(u, v)$, though in practice the signing of alleles has little impact on our results.

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

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

. When a, b is far from locations in Ω relative to the scale s, the wavelet function ψ is only evaluated over a range where it remains close to zero, resulting in very small $h_{a,b}(s)$. In turn, small $h_{a,b}(s)$ inflates the wavelet transformed allele frequencies (equation 3), which is undesirable for small s in these geographic regions where there is little sampling. For this reason we do not calculate the wavelet transform for locations a, b where there are no locations sampled closer than $2s$ distance units.

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.

2.1.1 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. Here, we specifically want to characterize heterogeneity in spatial patterns, because this heterogeneity in pattern may reflect heterogeneity in underlying processes: where there is heterogeneity in migration rates, such as where there are migration barriers (Petkova, Novembre, and Stephens 2016), or where there are recent range expansions such that spatial patterns are farther from equilibrium (Slatkin 1993). For these reasons, the method presented here is aimed at characterizing localized and scale-specific patterns of isolation by distance.

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 (see simulations below, Figure 1E and 1F). A benefit of using the wavelet transformation 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 analyst's choosing. To obtain 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.

2.1.2 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 is 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) (Keitt 2007).

2.1.3 Simulated neutral patterns across a continuous landscape

To demonstrate the wavelet transformation of allele frequencies, and wavelet genetic dissimilarity function, we applied these tools to several simulated scenarios. First, we conducted forward landscape genetic simulations under neutrality using the SLiM software (Haller and Messer 2019), building off published approaches (Battey, Ralph, and Kern 2020). We simulated outcrossing, iteroparous, hermaphroditic organisms, with modest lifespans (average of $\sim$ 4 yrs/time steps). Mating probability was determined based on a Gaussian kernel as was dispersal distance from mother (Battey, Ralph, and Kern 2020). Individuals became mature in the time step following their dispersal. These parameters roughly approximate a short lived perennial plant with gene flow via pollen movement and seed dispersal. Code is available at GitHub (https://github.com/jesserlasky/WaveletSpatialGenetic).

We began by characterizing a simple scenario across a continuous landscape. We simulated a square two dimensional landscape measuring 25 units on each side. The standard deviation of mating and dispersal distance were both 0.2, yielding a combined standard deviation of gene flow distances of 0.28. In this first simulation there was no selection. 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 predict the spatial location of each sample (Figure S2).

To facilitate interpretation of wavelet transformed allele frequencies $(T^{wav} f_i)(a, b, s)$ we provide two example loci i with distinct spatial patterns (Figure 1). The first locus has the greatest variance in wavelet transformed allele frequencies among sampled loci at $s = 0.4$ (Figure 1A-C) while the second locus has the greatest variance at $s = 12.2$ (Figure 1D-F).

We then calculated wavelet dissimilarity $D^{wav}_{a,b}(s)$, aggregating the signals in $(T^{wav} f_i)(a, b, s)$ across loci i, 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, as genetic distances are often linearly correlated to log geographic distances in two dimensions (Rousset 1997). 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

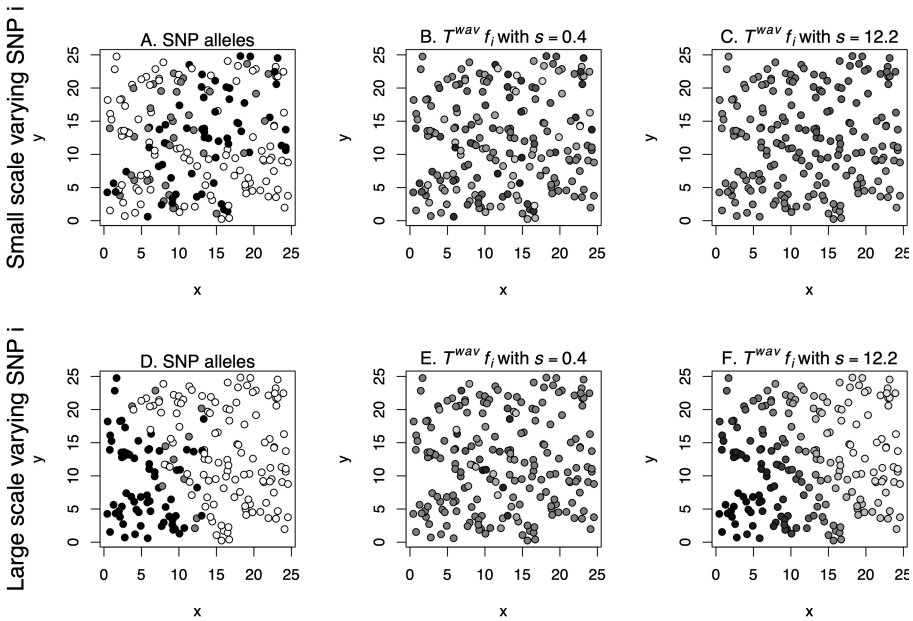

Figure 1: **Two example SNPs (rows) with distinct spatial patterns.** Shading shows either allelic variation (untransformed, A, D) or variation in wavelet transformed allele frequencies $(T^{wav} f_i)(a, b, s)$ (B,C,E,F). The first locus (A-C) has the greatest variance in wavelet transformed allele frequency among sampled loci at $s = 0.4$. The second locus (D-F) has the greatest variance in wavelet transformed allele frequency at $s = 12.2$. For the SNP in the top row, the variance among locations in $(T^{wav} f_i)(a, b, s)$ for $s = 0.4$ is 0.56 (visualized as shading in B), while it is only 0.17 for the SNP in the bottom row (E). For the SNP in the bottom row, the variance among locations in $(T^{wav} f_i)(a, b, s)$ for $s = 12.2$ is 44.46 (visualized as shading in F), while it is only 1.24 for the SNP in the top row (C).

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 (Figure S2).

To demonstrate how the scale of gene flow influences the wavelet dissimilarity $D^{wav}_{a,b}(s)$, we also conducted identical simulations as described above but instead with standard deviations of mating and dispersal distance of 0.5, 1, 2, or 5, yielding combined standard deviations of gene flow distances of 0.71, 1.41, 2.83, and 7.07. With increasing scale of gene flow we see a flatter change in wavelet dissimilarity across spatial scales (Figure S3). When gene flow is local, wavelet dissimilarity is low at small scales and high at large scales. At the large gene flow scale, the observed wavelet dissimilarity is indistinguishable from the panmictic null. These results give some intuition into how our approach characterizes spatial pattern in allele frequency, but this homogeneous system is not our target use case.

2.1.4 Simulated long-term neutral patterns in a heterogeneous landscape

To assess if our approach could identify localized and scale-specific patterns of isolation by distance, we next simulated multiple scenarios where we expected greater spatial heterogeneity. First, we simulated neutral evolution across a patchy, heterogeneous landscape. We used simulated patchy landscapes (generated from earlier work) (Jesse R. Lasky and Keitt 2013). 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.

Additionally, we sought to connect the patterns identified by wavelet dissimilarities to those more familiar to many readers, so as to help readers gain intuition into our method. To do so, we calculated pairwise euclidean genetic and geographic distances between samples, and did this for different subsets of samples and regions, so as to connect localized patterns in wavelet dissimilarity to localized patterns in pairwise distances.

In our simulated heterogeneous landscape, wavelet dissimilarity showed localized and scale-specific patterns of low and high dissimilarity (Figure 2). Notably, the same two islands (top left and bottom right of landscape in Figure 2) have lower dissimilarity than expected at small scales and are more dissimilar than expected at larger scales. Stated another way, these islands have low diversity locally (e.g. within populations), as can be seen by the slow increase

in genetic distance with geographic distance locally (Figure 2D, compare to 2F). However, at larger scales (e.g. comparing island to mainland) islands are more dissimilar, as seen by the greater genetic distances at larger geographic distances (Figure 2E, compare to 2G; also see the first two principal components of SNPs, Figure S4). These results highlight the capacity of the method to contrast patterns across scales and only requiring dilation of the analyzing kernel.

2.1.5 Simulated neutral patterns in a colonizing and range-expanding species

For a second scenario where we expected localized, scale-specific heterogeneity in spatial population structure, we simulated an invasion/range expansion. Beyond the importance of invasions in applied biology, the changes in spatial genetic patterns over time are of general interest (Castric and Bernatchez 2003; Le Corre et al. 1997; Slatkin 1991, 1993), considering that all species ranges are dynamic and many "native" species still bear clear evidence of expansion, e.g. following the last glacial maximum.

We simulated invasion across a square landscape of the same size as before, but beginning with identical individuals only in the middle at the bottom edge of the landscape (Figure 3). 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 heterogeneity in spatial patterns in allele frequency, demonstrated by the variation in wavelet dissimilarity across different regions and across time (e.g. for $s = 6.9$, dark versus light red in Figure 3A-C). This heterogeneity in genetic turnover 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 3D). In the range interior, there is greater local diversity and less turnover in genotype across space, i.e. a weaker isolation by distance (Figure 3E, see all SNP genetic distance plot Figure S4). 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 (Figure 3G) after the landscape had been populated.

These patterns highlight how wavelet dissimilarity is capturing scale-specific turnover in genetic composition, rather than merely genetic distance at a given geographic distance. Comparing the two regions highlighted in Figure 3B, the genetic distances at a geographic distance of 6.9 are not strikingly different (Figure S5). Rather what distinguishes these regions is their rate of genetic change in composition at this scale, as highlighted in Figure 3. The region of high wavelet dissimilarity at $s = 6.9$ (Figure 3B) transitions from homogeneity among nearby samples to high genetic distance at larger scales (Figure 3D, S5). By contrast the region of low wavelet dissimilarity at $s = 6.9$ (Figure 3B) starts

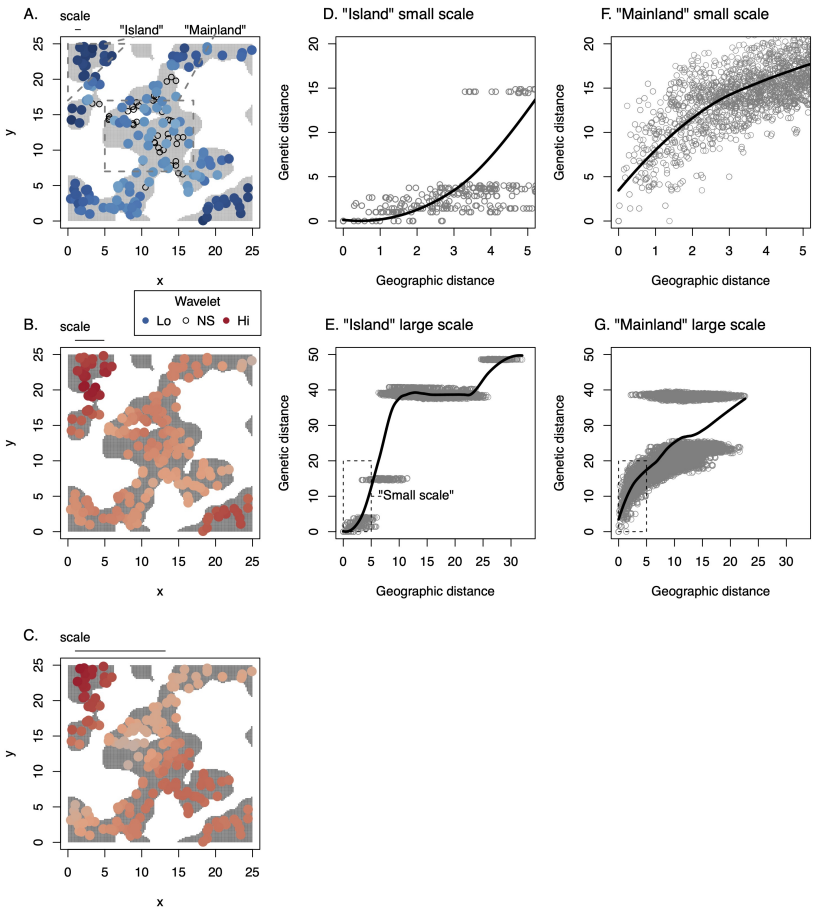

Figure 2: **Wavelet genetic dissimilarity identifies scale-specific, localized patterns in a heterogeneous landscape, with pairwise distance plots for comparison.** (A-C) Maps of simulated landscape where habitat is gray (in background) and unsuitable areas are white. Sampled individuals are circles. 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 as a horizontal line. At the smallest scales (A), samples have less dissimilarity than expected, especially in the island in the upper left of the landscape. This pattern can also be seen (D,F) when comparing pairwise geographic versus genetic distances for samples in the different regions of the landscape (dashed grey lines in A). At larger spatial scales (B-C), all locations have significantly greater dissimilarity than expected due to limited gene flow. However, the same islands show the greatest dissimilarity at large scales (lower panels), due to their high genetic difference from mainland samples at center. This pattern can also be seen in the pairwise genetic distances across larger geographic distances (E,G). (D-G) Loess smoothing curves are shown.

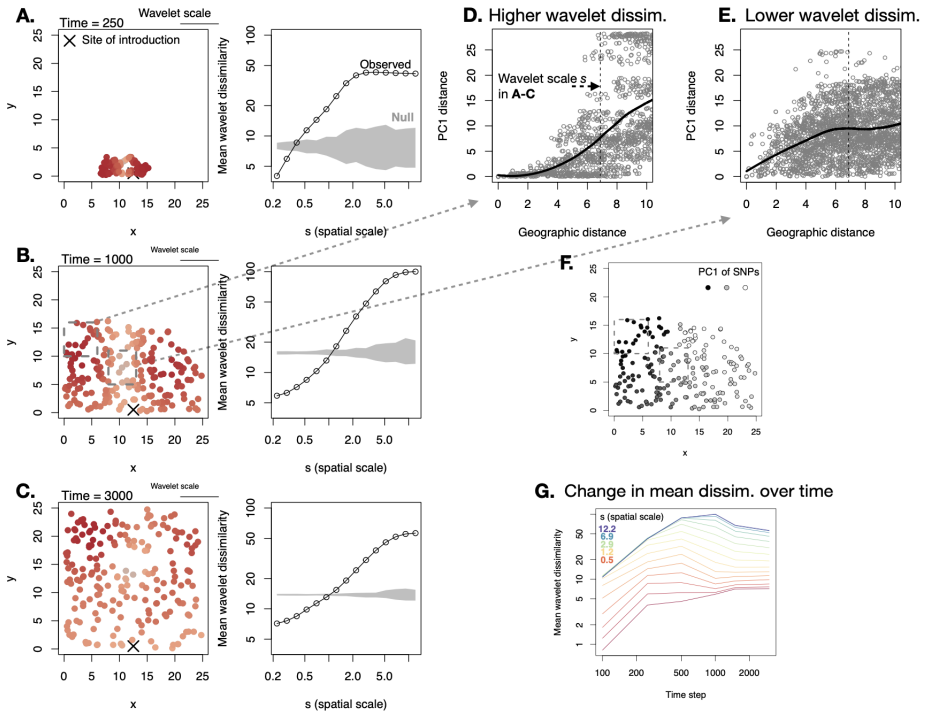

Figure 3: **Wavelet genetic dissimilarity reveals dynamic spatial patterns during an invasion across a homogeneous landscape.** Left column of panels (A-C) shows a map of the landscape through time, with 200 sampled individuals at each time step and the wavelet dissimilarity at $s = 6.9$ at their location. Darker red indicates greater wavelet dissimilarity. In the second time step, 1000, two regions are highlighted in dashed boxes (B), one with higher dissimilarity at $s = 6.9$ (D) and one with lower dissimilarity at this scale (E). (D-E) show pairwise geographic distance versus distance in the first PC of SNPs for samples from these regions. (F) shows the loadings of each sample on the first PC of SNPs. (D-E) highlight the greater increase in PC1 distance with geographic distance at this scale (vertical dashed lines) in (D), compared to the smaller increase in PC1 distance at this scale in (E). In particular, the region highlighted in (D) is homogeneous at short distances but very distinct at distances at the highlighted scale $s = 6.9$, indicating the major genetic turnover at this scale and location. (G) Mean wavelet dissimilarity across the landscape changes over time, highlighting the dynamic spatial population genetic patterns across invasions. Loess smoothing curves are shown in (E-F).

out with greater genetic distance among nearby samples with a modest increase in genetic distance at larger scales (Figure 3E, SX).

Overall, these simulations show the capacity of $D^{wav}_{a,b}(s)$, wavelet genetic dissimilarity, to capture localized, scale specific trends in genetic composition. Given the spatial heterogeneity in nature and the dynamics of populations and species ranges through time, there are likely many such patterns waiting to be described to shed light on patterns of gene flow and population history.

2.2 Finding the loci of local adaptation

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

We can also use our approach to transforming allele frequencies to identify particular genetic loci involved in local adaptation, and the regions and spatial scales of turnover in their allele frequency. Our strategy is (as before) to first calculate $(T^{wav} f_i)(a, b, s)$, the wavelet transform, for each locus i at each sampling point a, b for a set of chosen spatial scales $s \in S$.

Because of different ages and histories of drift, mutations will vary in their global allele frequency and thus global variance. To facilitate comparisons among loci for relative evidence of selection, we can normalize spatial patterns in allele frequency by total variation across locations, as is done when calculating F_{ST}.

Here we divide the wavelet transforms of allele frequency by the standard deviation of global allele frequency variation for each locus i, $sd(f_i)$. This normalization is greatest when minor allele frequency is 0.5 for a biallelic locus, and yields a scaled wavelet transformed 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 equivalent standard deviation and variance equal to unity.

We then calculate the variance across sampling locations of $(T^{wav} f_i)(a, b, s)/sd(f_i)$ and refer to this quantity as the "scale-specific genetic variance." This scaled-specific 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 scale-specific 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 (Cavalli-Sforza 1966; Lewontin and Krakauer 1973) and used the approach of Whitlock and Lotterhos (2015) to fit the degrees of freedom of this distribution to the distribution of scale-specific genetic variances (see Supplemental Methods). Applying this approach to a range of simulated scenarios as well as an empirical dataset (described below), we see that the χ^2 distribution with a maximum likelihood fit to determine degrees of freedom provides a reasonably close fit to the distribution of scale-specific genetic variance among SNPs (Figures S6-S9).

2.2.2 Simulated local adaptation

First, we present some specific individual simulations for illustration, and then a larger set with more variation in underlying parameters. 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. As with the neutral simulations, simulations with selection 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. Full details of simulation, including complete code, can be found in supplemental materials and on GitHub (https://github.com/jesserlasky/WaveletSpatialGenetic).

In the first simulation along the linear gradient, there were 2 selected loci with minor allele frequency (MAF) at least 0.1, 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 the scale-specific genetic variance $var((T^{wav} f_i)(a, b, s)/sd(f_i))$ at the larger spatial scales (Figure 4). 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 scale-specific genetic variance was at the largest scale (Figure 4). 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 4). 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 change in selection in a discrete habitat patch, which may more closely correspond to the setting where researchers would find useful a flexible approach to finding spatial patterns in allele frequency, especially if the patches of distinct environment are not known by researchers. In our simulation there was a large central patch, 10x10 units, 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 diversity, 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 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

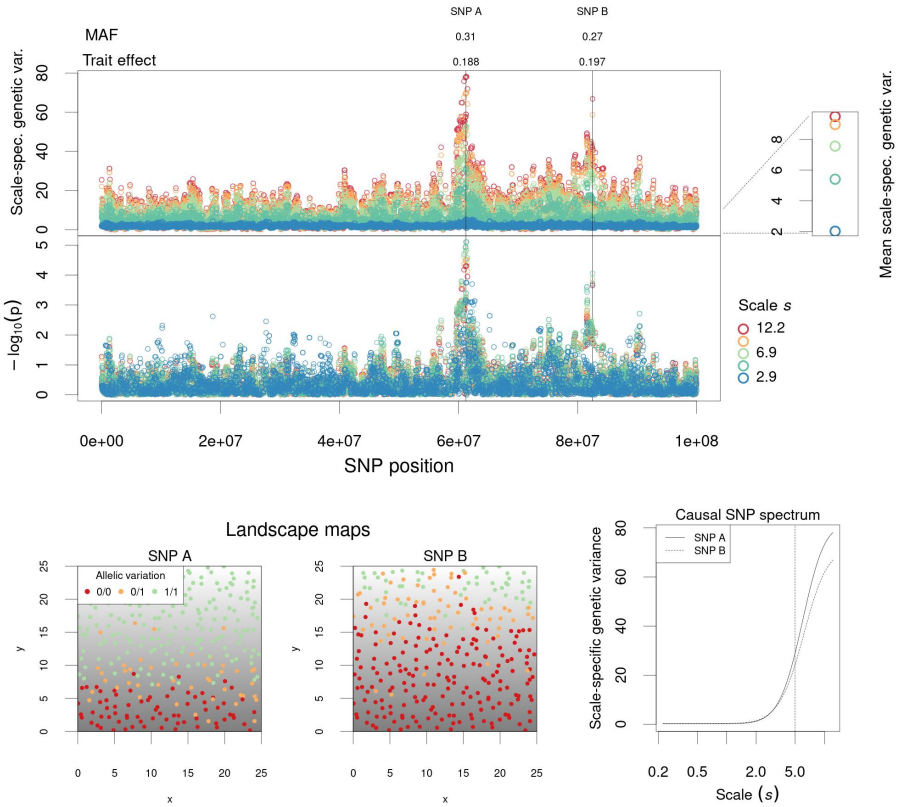

Figure 4: **Scale-specific genetic variance test applied to simulations with a linear selective gradient**. (top panels) Genome-wide variation in scale-specific genetic variance, $var((T^{wav} f_i)(a, b, s)/sd(f_i))$, for five different scales s and upper-tail p-values for χ^2 test using fitted values of d.f. Each point represents a SNP at a specific scale. 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 bottom are shown maps of the two selected loci as well as their spectra of scale-specific genetic variance. At upper right the mean scale-specific genetic 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 approximately 5 individuals per square unit area.

peaks in scale-specific genetic variance (Figure 5A), but when using the χ^2 test on these SNP variances we clearly identified the single QTL under selection, with lowest p-values for intermediate scales (Figure 5B).

We calculated the scale-specific genetic variance across a denser spectrum of scales s for the causal SNP, to determine at what scale variance was greatest. We found the maximum scale-specific genetic variance for the causal SNP was at 5.02, approximately half the length of a patch edge (Figure 5C). For illustration, we also calculated F_{ST} (Goudet 2005; Weir and Cockerham 1984) for two naively discretized subpopulation scenarios to show how this discretization can go awry (Figure 5D-F).

2.2.3 Evaluating the scale-specific genetic variance test

As an initial assessment of the general appropriateness of the scale-specific genetic variance test we proposed above, we conducted additional simulations on two types of landscapes with varying life history parameters. 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 discrete habitat patch landscape and the linear gradient landscape but with a wider range of parameter variation. We tested a range of mating and dispersal (σ) scales including 0.25, 0.5, 1, and 2, and a range of stabilizing selection (σ_k) values including 0.125, 0.25, 0.5, and 1. Three simulations were conducted for each combination of parameter settings and each ran for 10,000 years. Because PCAdapt is one of the few methods for identification of spatial pattern in allele frequency that does not require subpopulation discretization and in theory could detect patterns at multiple scales, we also implemented this method. We used the PCA of the scaled genotype matrix, thinned for LD but including causal SNPs, to extract the z-scores and p-values of each SNP with a cutoff of p = 0.05. We used a scree plot showing the percentage of variance explained in decreasing order to identify the optimal number of principal components following Cattell's rule (Duforet-Frebourg et al. 2016).

For the scale-specific genetic variance test results, we focus on the scale we tested most closely corresponding to the scale of the patch, $s = 5.2$ (the patch was a 10x10 square). Overall the scale-specific genetic variance test 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 (Figure 6). By contrast, under scenarios of low gene flow and strong stabilizing selection, nominal false positive rates were high for PCAdapt, often > 0.15.

Power to detect SNPs (proportion of selected SNPs with $p < 0.05$) under selection was generally high (true positive rate near 1) but sometimes low, depending on the strength of selection (σ_k) and mating and dispersal scales (σ) (Figure 6). When gene flow was high and selection was weak, power was low for both the scale-specific genetic variance test and PCAdapt. This also corresponds to the scenario when local adaptation is weakest (Kirkpatrick and Barton

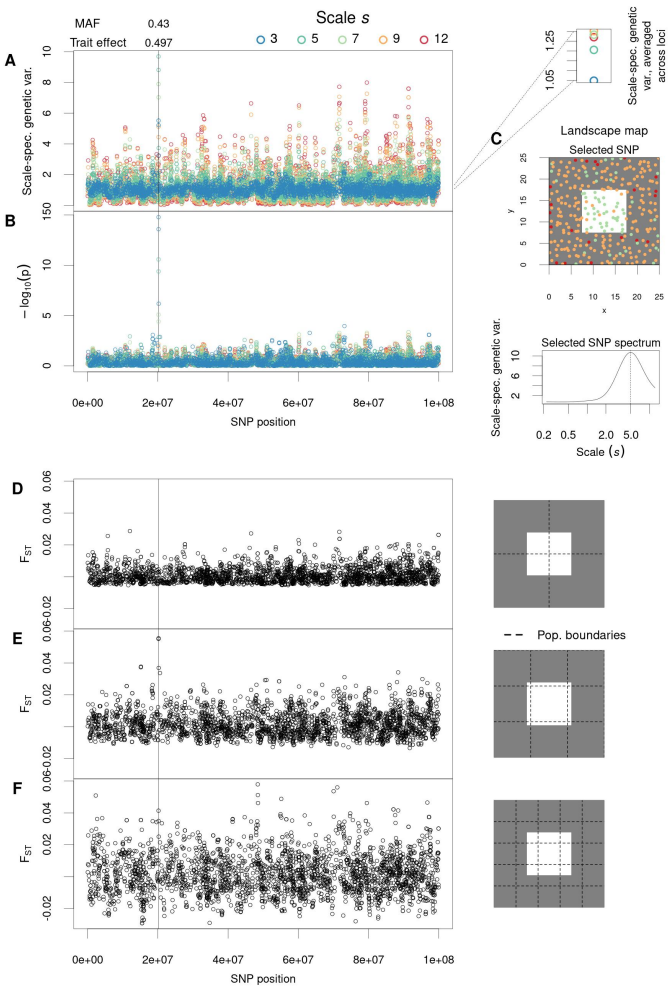

Figure 5: **Simulations of local adaptation to a single discrete patch of different habitat**. (A) Genome-wide variation in scale-specific genetic variance $var((T^{wav} f_i)(a, b, s)/sd(f_i))$ and (B) χ^2 p-values for six different scales s, for a discrete habitat difference after 3000 simulated years. Each point in the left panels represents a SNP, and wavelet statistics (A-B) at specific scales. The selected SNP is indicated with a vertical line along with the absolute value of a derived allele's effect on the trait and MAF. (C) A map of the landscape with individuals' genotypes at the causal SNP indicated with color, in addition to the spectrum of scale-specific genetic variance at this SNP, showing a peak at approximately half the patch width (vertical line at 5).(D-E) Implementation of F_{ST} using arbitrary boundaries for populations. This approach can easily miss causal loci (C,E) if the delineated population boundaries do not match habitat boundaries. (A) At upper right the mean scale-specific genetic 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 $\sigma_k = 0.08$.

1997). In addition to considering power simply based on p for each SNP, we also considered power using the top p-value rank among selected SNPs under each simulation, based on the reasoning that researchers may want to follow up on top ranked outlier SNPs first before any lower ranked SNPs. This approach showed similar results, with high power for both the scale-specific genetic variance test and PCAdapt except when gene flow was high and selection weak. In general, the two methods showed comparable power across different scenarios (Figure 6), with some indication that the scale-specific genetic variance test had higher power under high gene flow and PCAdapt slightly higher power under lower gene flow. By plotting individual SNPs we can see that for the upper end of gene flow scenarios ($\sigma = 1$ or 2), the scale-specific genetic variance test more consistently identified selected SNPs at the top compared to PCAdapt. For the low gene flow scenarios, PCAdapt more consistently identified large effect variants, while the scale-specific genetic variance test more consistently identified the smaller effect variants (see results for linear gradient in Figure S10). Overall, the similarities in true and false positive rates between methods suggest that our wavelet approach is effective compared to other related tools, while our test also offers the ability to explicitly consider variation in spatial scale.

2.3 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 (Berg and Coop 2014; Price et al. 2018). 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 calculate the mean of scale-specific genetic variance for all QTL with MAF at least 0.05 among sampled individuals. We then permute the identity of causal QTL across the genome and recalculate the mean scale-specific genetic variance, and repeat this process 1000 times to generate a null distribution of mean scale-specific genetic variance of QTL for each scale s.

We illustrate this test here briefly using a simulation of adaptation to a square patch of habitat in the middle of a landscape, with the two gene flow parameters $\sigma = 0.5$, the strength of selection $\sigma_K = 0.5$, carrying capacity ~ 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 scale-specific genetic variance for these QTL across scales s and compared to the null permutations of randomly selected 13 SNPs from the genome.

We found significantly higher mean scale-specific genetic variance for the QTL than the null expectation at all 6 scales tested. Although the scale-specific genetic variance was 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

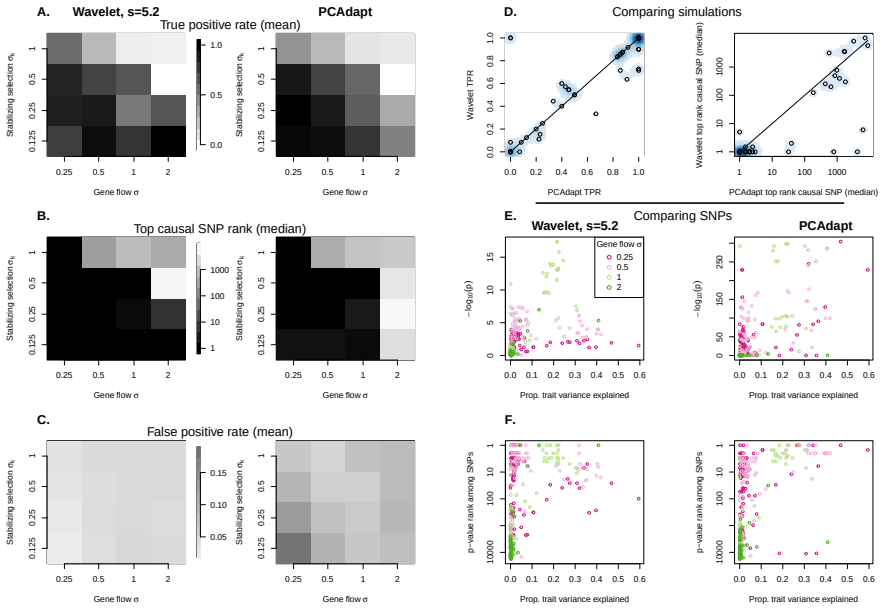

Figure 6: **Comparing the scale-specific genetic variance test with PCAdapt in simulations of adaptation to single discrete patch of different habitat**. (A) True positive rates (nominal $p < 0.05$) for each combination of simulation parameters, the scales of mating and dispersal σ and the standard deviation of the Gaussian stabilizing selection function σ_k. (B) An alternate view of statistical power based on the median rank of the top selected SNP among all SNPs. (C) False positive rates (nominal $p < 0.05$). (D) Comparing power between the two statistical approaches for the different simulation runs. Density of points is shown in the blue scale so as to indicate where many simulations had the same result. The line indiciates a 1:1 relationship. (E-F) Individual selected SNPs in simulations, showing their nominal p values and ranks among all SNPs, colored based on σ in the simulation. The x-axis represents the proportion of total phenotypic variation among sampled individuals that was explained by the given SNP (R^2 from a linear model).

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 S11).

2.4 Application to an empirical system

2.4.1 Genome-wide wavelet dissimilarity

We applied our approach to an empirical dataset of diverse, broadly distributed genotypes with whole genome resequencing data: 908 genotypes from 680 natural populations of the model plant, *Arabidopsis thaliana* (Brassicaceae). We used a published Arabidopsis dataset (Alonso-Blanco et al. 2016), only including Eurasian populations and excluding highly distinct "relicts" and also likely contaminant accessions (Pisupati et al. 2017). For locations with more than one accession genotyped we calculated allele frequency. We used a total of 129536 SNPs filtered for minor allele frequency (MAF$>$ 0.05) and LD (Zheng et al. 2012).

We first calculated the genome-wide wavelet dissimilarity, $D^{wav}_{a,b}(s)$, across a series of increasing scales s at even intervals in log distance units from $\sim$ 50 m to approximately half the distance separating the farthest samples, $\sim$ 3000 km. When samples had no other samples within a distance of $2s$ we excluded those samples from the dissimilarity calculation at that scale s.

We observed increasing mean genome-wide wavelet dissimilarity at larger scales (Figure 7), a pattern indicative of isolation by distance, on average, across the landscape. Arabidopsis showed significantly low dissimilarity at scales less than $\sim$ 5 km, likely due to the homogenizing effect of gene flow. However, we found significantly high dissimilarity at scales greater than $\sim$ 7 km. This scale of significantly high dissimilarity may be a relatively short distance, considering that Arabidopsis is largely self pollinating and lacks clear seed dispersal mechanisms (though seeds of some genotypes form mucus in water that increases buoyancy) (Saez-Aguayo et al. 2014). At scales greater than $\sim$ 120 km we found an increase in the slope relating scale s and dissimilarity, perhaps signifying a scale at which local adaptation begins to emerge.

The locations of scale-specific dissimilarity among Arabidopsis populations revealed several interesting patterns. Even by the $\sim$ 30 km scale, there were three notable regions of significantly high dissimilarity: northern Spain and extreme southern and northern Sweden (Figure 7). The high dissimilarity at this scale in northern Spain 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. (2013) previously found a particularly steep increase in isolation-by-distance. Alonso-Blanco et al. (2016) found that genetic distance was greatest among accessions from Southern Sweden at scales from $\sim 20-250$ compared to regions farther south. At larger, among-region scales, dissimilarity was significantly high across the range, with Iberia and northern Sweden again being most dissimilar at $\sim$ 234 km and surpassed by central Asia by $\sim$ 1834 km as being most dissimilar. Iberia and northern Sweden contain many accessions

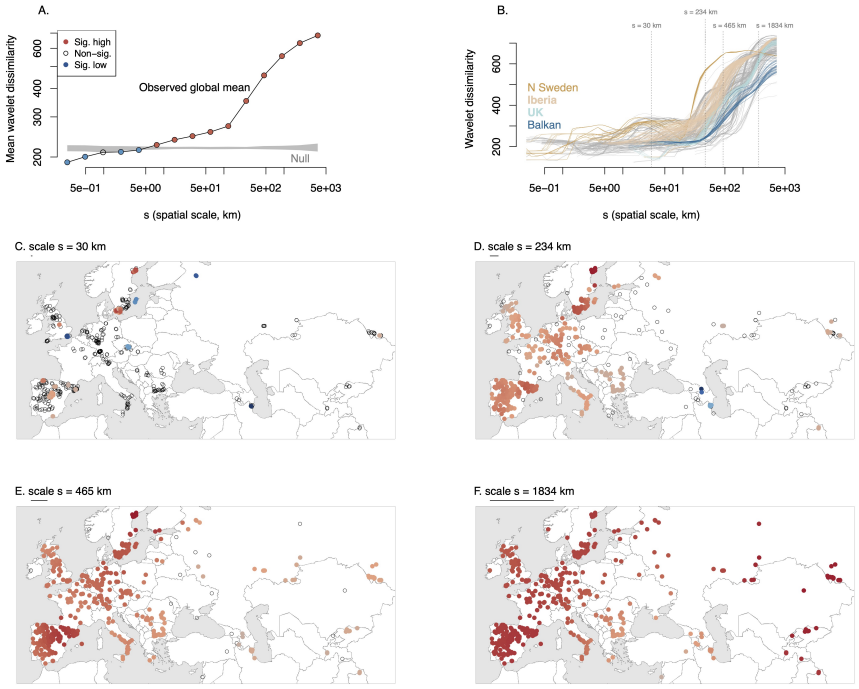

Figure 7: **Genome-wide wavelet dissimilarity, $D^{wav}_{a,b}(s)$, for Arabidopsis genotypes.** (A) The global mean dissimilarity across scales compared to the null expectation (gray ribbon) and (B) the dissimilarity across scales centered on each sampled genotype, with several regions highlighted (vertical lines indicate scales shown in panels C-F). (C-F) Selected scales highlight the changes in dissimilarity across locations, with each circle indicating a genotyped sample/population. 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 (for a given scale) in $D^{wav}_{a,b}(s)$ among significant locations.

distantly related to other accessions, likely due to isolation during glaciation and subsequent demographic histories (Alonso-Blanco et al. 2016). 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. By contrast, populations in the UK and the Balkan peninsula had low dissimilarity across a range of scales, possibly due to reduced diversity and a more recent history of spread in these regions.

2.4.2 Identifying putative locally-adapted loci

For this analysis, we used the same genotypes as in the prior section but not filtered for LD, leaving 1,642,040 SNPs with MAF$>$ 0.1 (Alonso-Blanco et al. 2016) .

The scale-specific genetic variance test identified putative locally adapted loci (Figure S12). The distribution of scale-specific genetic variance among SNPs was reasonably matched to the theoretical χ^2 distribution, with some deviation at low values of scale-specific genetic variance (Figure S9). Among notable loci, at the $\sim$ 59 km scale, the #2 QTl and #3 SNP is in the coding region of METACASPASE 4 (MC4), a gene that controls biotic and abiotic stress-induced programmed cell death (Hander et al. 2019; Shen, Liu, and Li 2019). To speculate, if MC4 were involved in coevolution with microbial pathogens we might expect a pattern of high variation among even nearby populations due to rapid allele frequency dynamics.

The #1 SNP for the $\sim$ 282 km scale was in the coding sequence of the DOG1 gene (Figure 8, Figure S12). This SNP, Chr. 5, 18,590,741 was also strongly associated with flowering time (see next section) and germination and tags known functional polymorphisms at this gene that are likely locally adaptive (Martínez-Berdeja et al. 2020). The spatial pattern of variation at this locus (Figure 8) 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 F_{ST} could easily miss the signal as did Horton et al. (2012). The climate-allele frequency associations for DOG1 are also complicated and non-monotonic (Gamba et al. 2022; Martínez-Berdeja et al. 2020), making it challenging for genotype-environment association approaches (Jesse R Lasky, Emily B Josephs, and Morris 2023).

At the $\sim$ 1359 km scale, the #1 SNP (and also the lowest p-value SNP among all scales, Figure 8, Figure S12) was on chromosome 5 at 26,247,515 bp, 555 bp upstream from AT5G65660, a hydroxyproline-rich glycoprotein family protein. These are cell wall glycoproteins with important roles in development and growth (Johnson et al. 2017) some of which have a role in abiotic stress response (Tseng et al. 2013).

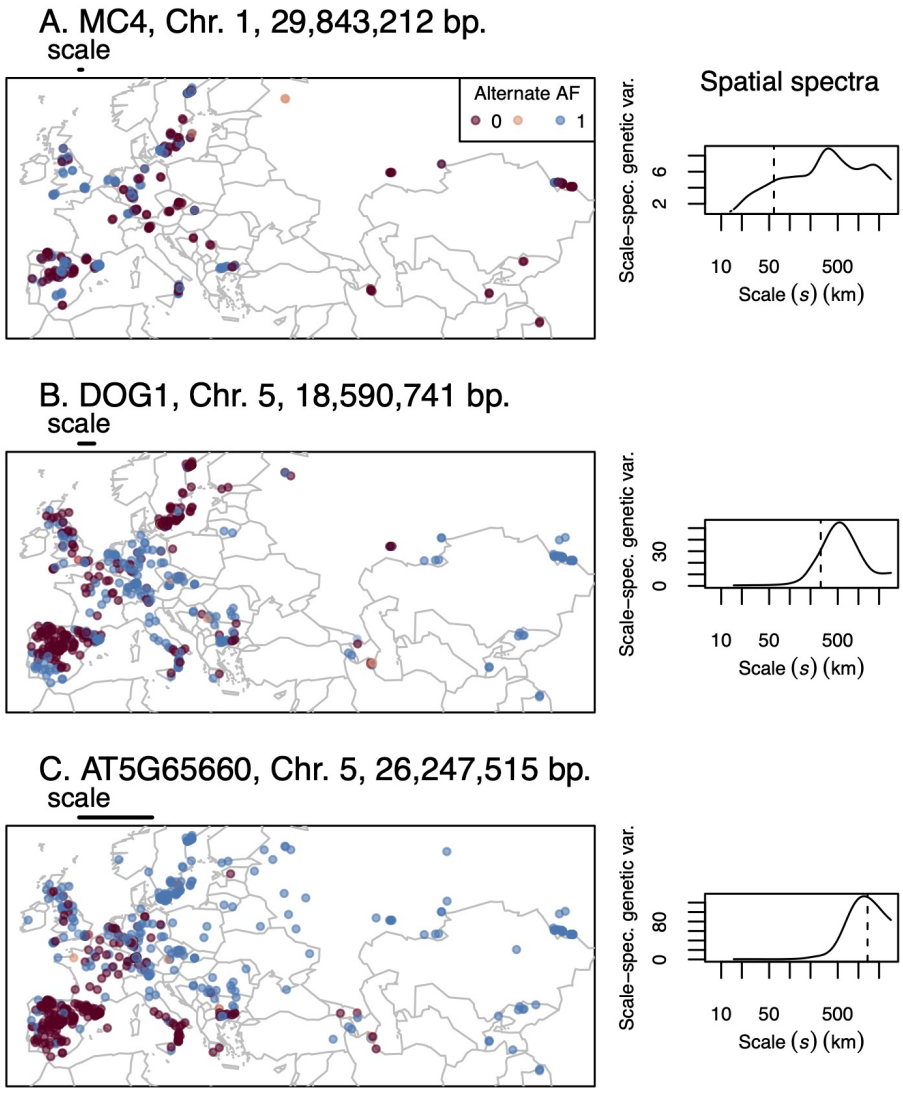

Figure 8: **Allelic variation (colors) for SNPs that were top outliers for scale-specific genetic variance test at different scales.** On maps at left, the scale for which a SNP was an outlier is indicated by a bar above each map. The right panels show the spatial spectra for each SNP, i.e. the scale-specific genetic variance across a range of scales. Dashed lines indicate the scale for which a SNP was an outlier.

2.4.3 Testing for local adaptation in quantitative trait loci (QTL)

We tested for non-random scale-specific genetic variance of QTL for Arabidopsis flowering time, a trait that is likely involved in local adaptation (Ågren et al. 2017). We used previously published data on flowering time: days to flower at 10°C measured on 1003 genotypes and days to flower at 16°C measured on 970 resequenced genotypes (Alonso-Blanco et al. 2016). We then performed mixed-model genome wide association studies (GWAS) in GEMMA (v 0.98.3) (Zhou and Stephens 2012) 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 scale-specific genetic variance at several intermediate spatial scales tested. For flowering time at both 10° and 16°C, scale-specific genetic variance was significantly elevated for the top 1,000 SNPs at the 282, 619, and 1359 km scales, but not always at the largest or smallest scales (Figure 9). In particular the scale-specific genetic variances were greatest for the $\sim$ 282 km scale where the mean scale specific genetic variance for 16°C QTL was 15.2 standard deviations above the null mean, and the $\sim$ 619 km scale, where the mean scale specific genetic variance for 10°C QTL was 13.5 standard deviations above the null mean. For QTL from both temperature experiments, results were nearly equivalent if we instead used the top 100 SNPs.

3 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 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 (Alonso-Blanco et al. 2016; Horton et al. 2012).

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 2, 3, 7). We also showed how wavelet transforms can capture scale-specific evidence of

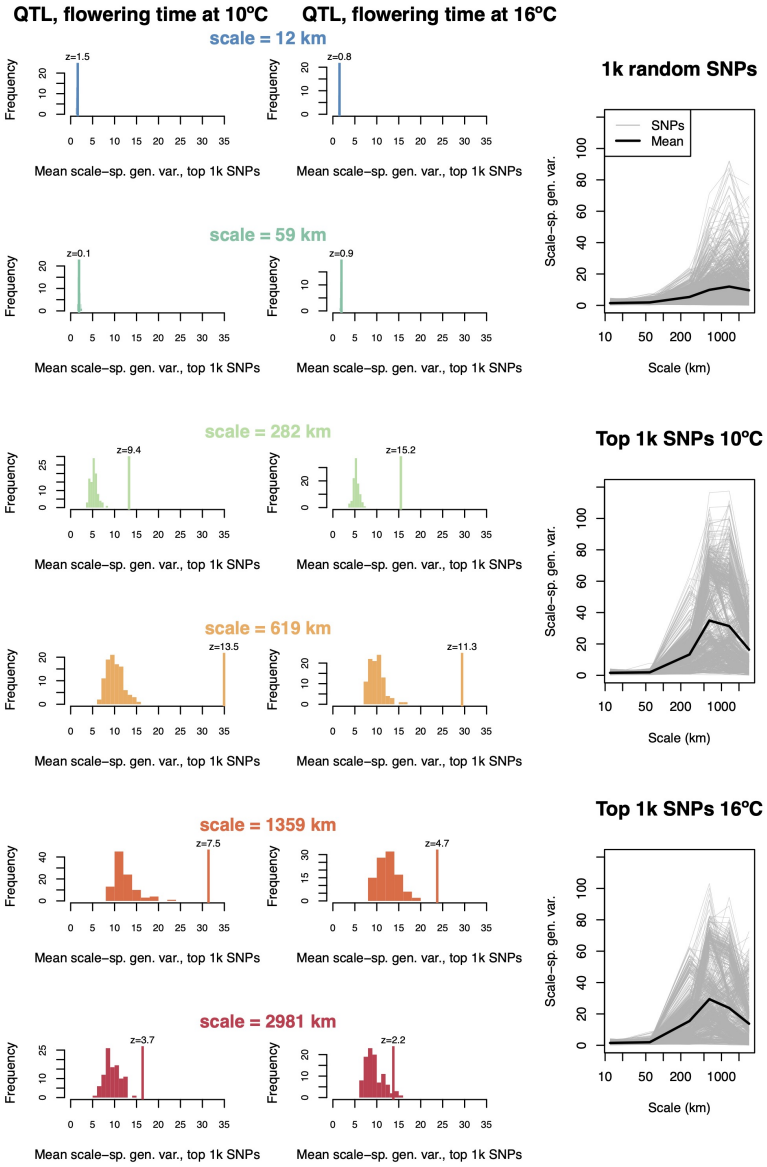

Figure 9: **Testing for selection on Arabidopsis flowering time QTL**. We compared scale-specific genetic variance, $var((T^{wav} f_i)(a, b, s)/sd(f_i))$, of QTL with random SNPs, for five different scales s, for flowering time measured at 10ºC and 16ºC. The first two columns show the observed mean of the top 1,000 flowering time SNPs with a vertical line and a z-score. The histograms show null distributions of scale-specific genetic variance based on permutations of an equal number of markers with an equal distribution as the flowering time QTL. At right the scale-specific genetic variance is shown for random SNPs and for the flowering time QTL (gray lines), across scales, with the mean indicated by a black line.

selection on individual genetic loci (Figures 4, 5, 6, 8) and on groups of quantitative trait loci (Figure 9). Our simulations and empirical examples showed substantial heterogeneity in the scale and stationarity 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 3). Additionally, we identified loci underlying local adaptation and showed an example where the evidence for this adaptation was specific to intermediate spatial scales (Figure 5). 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 (Wagner, Chávez-Pesqueira, and Forester 2017) or SNPs (Emily B. Josephs et al. 2019), and some can capture localized patterns (e.g Petkova, Novembre, and Stephens 2016), there are few examples of approaches that merge both abilities (Wagner, Chávez-Pesqueira, and Forester 2017).

The test for spatial pattern in individual loci we developed owes greatly to previous work from Lewontin and Krakauer (1973) who initially developed χ^2 tests applied to the distribution of F_{ST} values, and from Whitlock and Lotterhos (2015)'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 (François et al. 2016). However, we note that this test may be slightly conservative in some situations (Figure 6). Nevertheless, we believe there were important signs in our work that this χ^2-based scale-specific genetic variance test was valuable. In particular, we found in our simulation of adaptation to a habitat patch that the scale-specific genetic variance was greatest at large spatial scales but at neutral sites, which obscured spatial pattern at the causal locus (Figure 5). When applying the χ^2 test, we were able to clearly map the causal locus while spurious loci with high scale-specific genetic 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 9), even when the selective gradient was linear across the landscape (i.e. the largest scale, Figure 4). 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 versus higher relatedness between locations due to 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 (Heisenberg 1927). 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 (Emily B. Josephs et al. 2019; Whitlock and Guillaume 2009). Second, we could follow the example of Al-Asadi et al. (2019) 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 more 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.

3.1 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.

4 Materials and methods

4.1 Simulations with SLiM

We developed our simulations by building off the spatial neutral simulation model of Battey et al. Battey, Ralph, and Kern 2020 and model recipes in the SLiM software Haller and Messer 2019 for spatially varying selection on quantitative traits. Scripts are freely available at https://github.com/jesserlasky/WaveletSpatialGenetic/. Parameters differed among scenarios as described previously. The genome had 10^8 positions with a mutation rate of 10^{-7}. For the initial simulations to demonstrate patterns detectable by our test (Figures 4 and 5), the recombination rate was 10^{-8} per generation. For the simulations used to compare true positive and

false positive rates of the scale-specfic genetic variance test and PCAdapt (Figure 6), we increased the recombination rate to 10^{-7} per generation and thinned SNPs for LD (while retaining selected SNPs) so as to calculate meaningful SNP-level positive rates across the genome. The carrying capacity per unit square area was roughly 25 (see code). 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% background probability.

4.2 Null model test for wavelet transformed patterns at individual loci

Even under neutrality, individual loci differ in their history and thus not all have identical spatial patterns. To develop a null expectation for the distribution of scale-specific genetic variance across loci, we use the basic approach of Cavalli-Sforza (1966) and Lewontin and Krakauer (1973). Lewontin and Krakauer (1973) 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 (Lewontin and Krakauer 1973). 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 (Lewontin and Krakauer 1975; Nei and Maruyama 1975; Robertson 1975; Whitlock and Lotterhos 2015).

To solve this problem of non-independence among populations we use the same strategy that Whitlock and Lotterhos (2015) applied to F_{ST}: we use the distribution of scale-specific genetic 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 Whitlock and Lotterhos (2015) method: we trimmed outliers (here the bottom 2.5% SNPs for scale-specific genetic variance) in scale-specific genetic 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 (2015)), recalculated outliers, and then refit the χ^2 distribution iteratively. Mean scale-specific genetic variance was also calculated in this process while excluding SNPs in the bottom 2.5% tail as well as those with significantly high scale-specific genetic 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 scale-specific genetic 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 Whitlock and Lotterhos 2015) 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.

5 Supporting information

6 Acknowledgments

This work was supported by NIH award R35GM138300 to JRL. We thank Emily Josephs and Benjamin Peter for helpful comments.

References

Ågren, Jon et al. (Mar. 1, 2017). "Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping". In: *Evolution* 71.3, pp. 550–564. ISSN: 0014-3820. DOI: 10.1111/evo.13126. URL: https://doi.org/10.1111/evo.13126 (visited on 08/28/2023).

Alonso-Blanco, Carlos et al. (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166, pp. 481–491. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.05.063.

Al-Asadi, Hussein et al. (Jan. 14, 2019). "Estimating recent migration and population-size surfaces". In: *PLOS Genetics* 15.1. Publisher: Public Library of Science, e1007908. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007908. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007908 (visited on 02/24/2022).

Battey, CJ, Peter L Ralph, and Andrew D Kern (June 8, 2020). "Predicting geographic location from genetic variation with deep neural networks". In: *eLife* 9. Ed. by Patricia J Wittkopp et al. Publisher: eLife Sciences Publications, Ltd, e54507. ISSN: 2050-084X. DOI: 10.7554/eLife.54507. URL: https://doi.org/10.7554/eLife.54507 (visited on 01/25/2022).

Berg, Jeremy J. and Graham Coop (Aug. 7, 2014). "A Population Genetic Signal of Polygenic Adaptation". In: *PLoS Genet* 10.8, e1004412. DOI: 10.1371/journal.pgen.1004412. URL: http://dx.doi.org/10.1371%2Fjournal.pgen.1004412.

Bhatia, Gaurav et al. (Sept. 1, 2013). "Estimating and interpreting FST: The impact of rare variants". In: *Genome Research* 23.9. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1514–1521. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.154831.113. URL: http://genome.cshlp.org/content/23/9/1514 (visited on 08/06/2020).

Castric, Vincent and Louis Bernatchez (Mar. 2003). "The rise and fall of isolation by distance in the anadromous brook charr (Salvelinus fontinalis Mitchill)." In: *Genetics* 163.3, pp. 983–996. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462472/ (visited on 11/28/2022).

Cavalli-Sforza, L. L. (Mar. 22, 1966). "Population structure and human evolution". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 164.995. Publisher: Royal Society, pp. 362–379. DOI: 10.1098/

rspb.1966.0038. URL: http://royalsocietypublishing-org/doi/10.1098/rspb.1966.0038 (visited on 02/10/2022).

Daubechies, Ingrid (1992). *Ten lectures on wavelets*. SIAM. ISBN: 0-89871-274-2.

Duforet-Frebourg, Nicolas et al. (Apr. 1, 2016). "Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data". In: *Molecular Biology and Evolution* 33.4, pp. 1082–1093. ISSN: 0737-4038. DOI: 10.1093/molbev/msv334. URL: https://doi.org/10.1093/molbev/msv334 (visited on 02/24/2022).

Excoffier, Laurent and Nicolas Ray (July 2008). "Surfing during population expansions promotes genetic revolutions and structuration". In: *Trends in Ecology & Evolution* 23.7, pp. 347–351. ISSN: 0169-5347. DOI: 10.1016/j.tree.2008.04.004.

Fitzpatrick, Matthew C. and Stephen R. Keller (2015). "Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation". In: *Ecology Letters* 18.1, pp. 1–16. ISSN: 1461-0248. DOI: 10.1111/ele.12376. URL: http://dx.doi.org/10.1111/ele.12376.

François, Olivier et al. (2016). "Controlling false discoveries in genome scans for selection". In: *Molecular Ecology* 25.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13513, pp. 454–469. ISSN: 1365-294X. DOI: 10.1111/mec.13513. URL: http://onlinelibrary-wiley-com/doi/abs/10.1111/mec.13513 (visited on 02/11/2022).

Gamba, Diana et al. (Mar. 25, 2022). *The genomics and physiology of abiotic stressors associated with global elevation gradients in Arabidopsis thaliana*. Section: New Results Type: article. bioRxiv, p. 2022.03.22.485410. DOI: 10.1101/2022.03.22.485410. URL: https://www.biorxiv.org/content/10.1101/2022.03.22.485410v1 (visited on 03/26/2022).

Gautier, Mathieu (Dec. 2015). "Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates". In: *Genetics* 201.4, pp. 1555–1579. ISSN: 0016-6731. DOI: 10.1534/genetics.115.181453. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676524/ (visited on 07/07/2021).

Goudet, Jérôme (2005). "hierfstat, a package for r to compute and test hierarchical F-statistics". In: *Molecular Ecology Notes* 5.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.8286.2004.00828.x, pp. 184–186. ISSN: 1471-8286. DOI: 10.1111/j.1471-8286.2004.00828.x. URL: http://onlinelibrary-wiley-com/doi/abs/10.1111/j.1471-8286.2004.00828.x (visited on 02/10/2022).

Haller, Benjamin C. and Philipp W. Messer (2019). "SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model". In: *Molecular Biology and Evolution* 36.3, pp. 632–637. DOI: 10.1093/molbev/msy228. URL: http://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msy228/5229931 (visited on 01/19/2019).

Hander, Tim et al. (Mar. 22, 2019). "Damage on plants activates Ca2+-dependent metacaspases for release of immunomodulatory peptides". In: *Science (New York, N.Y.)* 363.6433, eaar7486. ISSN: 1095-9203. DOI: 10.1126/science.aar7486.

Heisenberg, W. (Mar. 1, 1927). "Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik". In: *Zeitschrift für Physik* 43.3. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer-Verlag, pp. 172–198. DOI: 10.1007/BF01397280. URL: https://link.springer.com/article/10.1007/BF01397280 (visited on 03/21/2022).

Horton, Matthew W et al. (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel". In: *Nat Genet* 44, pp. 212–216. ISSN: 1546-1718. DOI: 10.1038/ng.1042. URL: http://dx.doi.org/10.1038/ng.1042.

Johnson, Kim L. et al. (June 1, 2017). "Insights into the Evolution of Hydroxyproline-Rich Glycoproteins from 1000 Plant Transcriptomes". In: *Plant Physiology* 174.2, pp. 904–921. ISSN: 0032-0889. DOI: 10.1104/pp.17.00295. URL: https://doi.org/10.1104/pp.17.00295 (visited on 08/24/2023).

Josephs, Emily B. et al. (Mar. 1, 2019). "Detecting Adaptive Differentiation in Structured Populations with Genomic Data and Common Gardens". In: *Genetics* 211.3. Publisher: Genetics Section: Investigations, pp. 989–1004. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.118.301786. URL: https://www.genetics.org/content/211/3/989 (visited on 11/18/2020).

Kawecki, Tadeusz J. and Dieter Ebert (2004). "Conceptual issues in local adaptation". In: *Ecology Letters* 7.12, pp. 1225–1241. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2004.00684.x. URL: http://dx.doi.org/10.1111/j.1461-0248.2004.00684.x.

Keitt, Timothy H. (2007). "On the quantification of local variation in biodiversity scaling using wavelets". In: *Scaling Biodiversity*. Publisher: Cambridge University Press, pp. 168–80.

Kirkpatrick, Mark and N. H. Barton (July 1, 1997). "Evolution of a Species' Range". In: *The American Naturalist* 150.1. ArticleType: research-article / Full publication date: July 1997 / Copyright © 1997 The University of Chicago Press, pp. 1–23. ISSN: 00030147. URL: http://www.jstor.org/stable/10.1086/286054.

Lasky, Jesse R, Emily B Josephs, and Geoffrey P Morris (Jan. 1, 2023). "Genotype–environment associations to reveal the molecular basis of environmental adaptation". In: *The Plant Cell* 35.1, pp. 125–138. ISSN: 1040-4651. DOI: 10.1093/plcell/koac267. URL: https://doi.org/10.1093/plcell/koac267 (visited on 02/26/2023).

Lasky, Jesse R., DAVID L. Des Marais, et al. (2012). "Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate". In: *Molecular Ecology* 21.22, pp. 5512–5529. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2012.05709.x. URL: http://dx.doi.org/10.1111/j.1365-294X.2012.05709.x.

Lasky, Jesse R. and Timothy H. Keitt (Nov. 1, 2013). "Reserve Size and Fragmentation Alter Community Assembly, Diversity, and Dynamics." In: *The American Naturalist* 182.5, E142–E160. ISSN: 0003-0147. DOI: 10.1086/673205. URL: http://dx.doi.org/10.1086/673205 (visited on 02/22/2017).

Le Corre, Valérie et al. (Apr. 1997). "Colonization with long-distance seed dispersal and genetic structure of maternally inherited genes in forest trees: a simulation study". In: *Genetics Research* 69.2. Publisher: Cambridge University Press, pp. 117–125. ISSN: 1469-5073, 0016-6723. DOI: 10.1017/S0016672397002668. URL: https://www.cambridge.org/core/journals/genetics-research/article/colonization-with-longdistance-seed-dispersal-and-genetic-structure-of-maternally-inherited-genes-in-forest-trees-a-simulation-study/FFEDD8A5B195F1C01C99261FE49B6432 (visited on 11/28/2022).

Lewontin, R. C. and J. Krakauer (May 1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms". In: *Genetics* 74.1, pp. 175–195. ISSN: 0016-6731.

— (June 1975). "Testing the Heterogeneity of F Values". In: *Genetics* 80.2, pp. 397–398. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1213337/ (visited on 02/10/2022).

Long, Quan et al. (Aug. 2013). "Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden". In: *Nature Genetics* 45.8. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 8 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Population genetics Subject_term_id: population-genetics, pp. 884–890. ISSN: 1546-1718. DOI: 10.1038/ng.2678. URL: http://www-nature-com/articles/ng.2678 (visited on 01/16/2022).

Machado, Heather E et al. (June 22, 2021). "Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in Drosophila". In: *eLife* 10. Ed. by Magnus Nordborg, Patricia J Wittkopp, and Magnus Nordborg. Publisher: eLife Sciences Publications, Ltd, e67577. ISSN: 2050-084X. DOI: 10.7554/eLife.67577. URL: https://doi.org/10.7554/eLife.67577 (visited on 01/24/2022).

Manel, Stéphanie et al. (Apr. 1, 2003). "Landscape genetics: combining landscape ecology and population genetics". In: *Trends in Ecology & Evolution* 18.4, pp. 189–197. ISSN: 0169-5347. DOI: 10.1016/S0169-5347(03)00008-9. URL: https://www.sciencedirect.com/science/article/pii/S0169534703000089 (visited on 10/16/2022).

Martínez-Berdeja, Alejandra et al. (Feb. 4, 2020). "Functional variants of DOG1 control seed chilling responses and variation in seasonal life-history strategies in Arabidopsis thaliana". In: *Proceedings of the National Academy of Sciences* 117.5. Publisher: National Academy of Sciences Section: Biological Sciences, pp. 2526–2534. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1912451117. URL: https://www.pnas.org/content/117/5/2526 (visited on 08/20/2021).

McRae, Brad H. et al. (2008). "Using Circuit Theory to Model Connectivity in Ecology, Evolution, and Conservation". In: *Ecology* 89.10. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-1861.1, pp. 2712–2724. ISSN: 1939-9170. DOI: 10.1890/07-1861.1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1890/07-1861.1 (visited on 08/27/2023).

McVean, Gil (Oct. 16, 2009). "A Genealogical Interpretation of Principal Components Analysis". In: *PLOS Genetics* 5.10. Publisher: Public Library of Science, e1000686. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000686. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686 (visited on 10/14/2022).

Muraki, S. (June 1995). "Multiscale volume representation by a DoG wavelet". In: *IEEE Transactions on Visualization and Computer Graphics* 1.2. Conference Name: IEEE Transactions on Visualization and Computer Graphics, pp. 109–116. ISSN: 1941-0506. DOI: 10.1109/2945.468408.

Nei, Masatoshi and Takeo Maruyama (June 1975). "Lewontin-Krakauertest for neutral genes". In: *Genetics* 80.2, p. 395. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1213335/ (visited on 02/10/2022).

Peter, Benjamin M, Desislava Petkova, and John Novembre (Apr. 1, 2020). "Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography". In: *Molecular Biology and Evolution* 37.4, pp. 943–951. ISSN: 0737-4038. DOI: 10.1093/molbev/msz280. URL: https://doi.org/10.1093/molbev/msz280 (visited on 01/09/2022).

Peterman, William E. (2018). "ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms". In: *Methods in Ecology and Evolution* 9.6, pp. 1638–1647. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12984. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12984 (visited on 04/14/2019).

Petkova, Desislava, John Novembre, and Matthew Stephens (Jan. 2016). "Visualizing spatial population structure with estimated effective migration surfaces". In: *Nature Genetics* 48.1. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Population genetics;Software Subject_term_id: population-genetics;software, pp. 94–100. ISSN: 1546-1718. DOI: 10.1038/ng.3464. URL: http://www-nature-com/articles/ng.3464 (visited on 09/10/2021).

Pisupati, Rahul et al. (Dec. 19, 2017). "Verification of Arabidopsis stock collections using SNPmatch, a tool for genotyping high-plexed samples". In: *Scientific Data* 4, p. 170184. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.184. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5744633/ (visited on 10/01/2018).

Price, Nicholas et al. (May 8, 2018). "Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana". In: *Proceedings of the National Academy of Sciences* 115.19, pp. 5028–5033. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1719998115. URL: http://www.pnas.org/content/115/19/5028 (visited on 06/06/2018).

Robertson, Alan (June 25, 1975). "Remarks on the Lewontin-Krakauer test". In: *Genetics* 80.2, p. 396. ISSN: 1943-2631. DOI: 10.1093/genetics/80.2.396. URL: https://doi.org/10.1093/genetics/80.2.396 (visited on 02/10/2022).

Rorschach, Hermann (1921). *Psychodiagnostik: Methodik und ergebnisse eines warhrnehmungsdiagnostischen Experiments (deutenlassen von zufallsformen)*. Vol. 2. E. Bircher.

Rousset, François (Apr. 1, 1997). "Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance". In: *Genetics* 145.4. Publisher: Oxford Academic, pp. 1219–1228. DOI: 10.1093/genetics/145.4.1219. URL: https://dx.doi.org/10.1093/genetics/145.4.1219 (visited on 08/27/2023).

Saez-Aguayo, Susana et al. (Mar. 13, 2014). "Local Evolution of Seed Flotation in Arabidopsis". In: *PLOS Genetics* 10.3. Publisher: Public Library of Science, e1004221. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004221. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004221 (visited on 02/13/2022).

Shen, Wenzhong, Jiuer Liu, and Jian-Feng Li (Nov. 4, 2019). "Type-II Metacaspases Mediate the Processing of Plant Elicitor Peptides in Arabidopsis". In: *Molecular Plant* 12.11, pp. 1524–1533. ISSN: 1674-2052. DOI: 10.1016/j.molp.2019.08.003. URL: https://www.sciencedirect.com/science/article/pii/S1674205219302643 (visited on 08/28/2023).

Slatkin, Montgomery (Oct. 1991). "Inbreeding coefficients and coalescence times". In: *Genetical Research* 58.2, pp. 167–175. ISSN: 0016-6723, 1469-5073. DOI: 10.1017/S0016672300029827. URL: https://www.cambridge.org/core/product/identifier/S0016672300029827/type/journal_article (visited on 11/28/2022).

— (1993). "Isolation by Distance in Equilibrium and Non-Equilibrium Populations". In: *Evolution* 47.1. Publisher: [Society for the Study of Evolution, Wiley], pp. 264–279. ISSN: 0014-3820. DOI: 10.2307/2410134. URL: http://www.jstor.org/stable/2410134 (visited on 10/16/2022).

Tseng, I-Chieh et al. (Sept. 1, 2013). "Abscisic Acid- and Stress-Induced Highly Proline-Rich Glycoproteins Regulate Root Growth in Rice". In: *Plant Physiology* 163.1, pp. 118–134. ISSN: 0032-0889. DOI: 10.1104/pp.113.217547. URL: https://doi.org/10.1104/pp.113.217547 (visited on 08/24/2023).

Wagner, Helene H., Mariana Chávez-Pesqueira, and Brenna R. Forester (2017). "Spatial detection of outlier loci with Moran eigenvector maps". In: *Molecular Ecology Resources* 17.6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12653, pp. 1122–1135. ISSN: 1755-0998. DOI: 10.1111/1755-0998.12653. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12653 (visited on 01/23/2022).

Wang, Ian J., Wesley K. Savage, and H. Bradley Shaffer (2009). "Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (Ambystoma californiense)". In: *Molecular Ecology* 18.7. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2009.04122.x, pp. 1365–1374. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2009.04122.x. URL: http://onlinelibrary-wiley-com/doi/abs/10.1111/j.1365-294X.2009.04122.x (visited on 10/16/2022).

Wang, Jianan et al. (Jan. 2020). "Genomic signatures of seed mass adaptation to global precipitation gradients in sorghum". In: *Heredity* 124.1. Bandiera_abtest:

a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Agricultural genetics;Evolutionary biology Subject_term_id: agricultural-genetics;evolutionary-biology, pp. 108–121. ISSN: 1365-2540. DOI: 10.1038/s41437-019-0249-4. URL: http://www-nature-com/articles/s41437-019-0249-4 (visited on 01/05/2022).

Weir, B. S. and C. Clark Cockerham (1984). "Estimating F-Statistics for the Analysis of Population Structure". In: *Evolution* 38.6. Publisher: [Society for the Study of Evolution, Wiley], pp. 1358–1370. ISSN: 0014-3820. DOI: 10.2307/2408641. URL: https://www.jstor.org/stable/2408641 (visited on 08/06/2020).

Whitlock, Michael C. and Frederic Guillaume (Nov. 2009). "Testing for Spatially Divergent Selection: Comparing QST to FST". In: *Genetics* 183.3, pp. 1055–1063. ISSN: 0016-6731. DOI: 10.1534/genetics.108.099812. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2778959/ (visited on 08/07/2020).

Whitlock, Michael C. and Katie E. Lotterhos (2015). "Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST". In: *The American Naturalist* 186 (S1). Publisher: [The University of Chicago Press, The American Society of Naturalists], S24–S36. ISSN: 0003-0147. DOI: 10.1086/682949. URL: https://www.jstor.org/stable/10.1086/682949 (visited on 08/30/2021).

Wright, Sewall (Jan. 1, 1949). "The genetical structure of populations". In: *Annals of Eugenics* 15.1. Publisher: John Wiley & Sons, Ltd, pp. 323–354. ISSN: 2050-1420. DOI: 10.1111/j.1469-1809.1949.tb02451.x. URL: http://onlinelibrary-wiley-com/doi/abs/10.1111/j.1469-1809.1949.tb02451.x (visited on 08/06/2020).

Yang, Wen-Yun et al. (June 2012). "A model-based approach for analysis of spatial structure in genetic data". In: *Nat Genet* 44.6, pp. 725–731. ISSN: 1061-4036. DOI: 10.1038/ng.2285. URL: http://dx.doi.org/10.1038/ng.2285.

Yeaman, Sam et al. (2016). "Convergent local adaptation to climate in distantly related conifers". In: *Science* 353.6306, pp. 1431–1433.

Zheng, Xiuwen et al. (Dec. 1, 2012). "A high-performance computing toolset for relatedness and principal component analysis of SNP data". In: *Bioinformatics* 28.24. Publisher: Oxford Academic, pp. 3326–3328. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts606. URL: http://academic.oup.com/bioinformatics/article/28/24/3326/245844 (visited on 04/13/2020).

Zhou, Xiang and Matthew Stephens (July 2012). "Genome-wide efficient mixed-model analysis for association studies". In: *Nat Genet* 44.7, pp. 821–824. ISSN: 1061-4036. DOI: 10.1038/ng.2310. URL: http://dx.doi.org/10.1038/ng.2310.

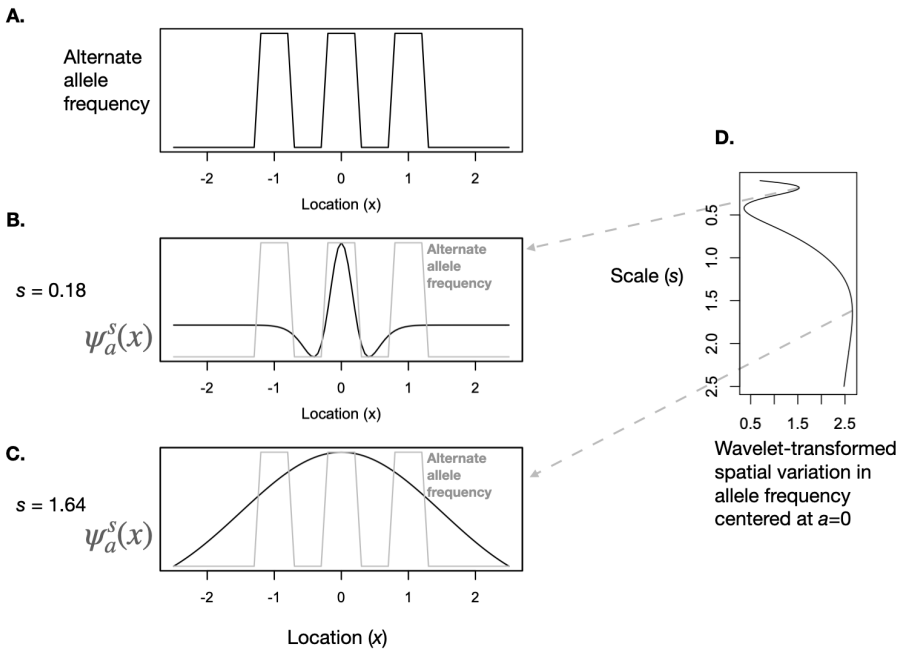

Fig. S1 An example of applying a difference of Gaussians (DoG) wavelet to spatial allele frequency patterns (here in one dimension). (A) shows the change in allele frequency across the spatial dimension x. (B-C) show two DoG wavelets (black curves) of two different scales s, centered at location $a = 0$, 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).

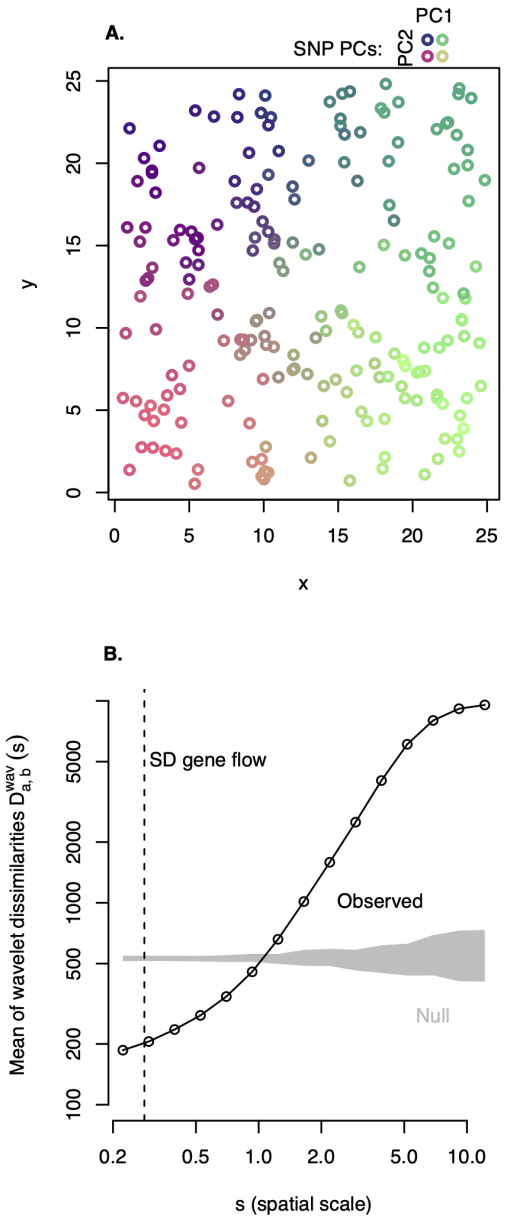

Fig. S2 Simulated two dimensional landscape. (A) with continuous population structure among 200 sampled individuals (circles), illustrated by the first two PCs of 1000 randomly selected SNPs (colors). In (A), each individual's color gives its SNP loadings on PC1 and PC2 according to the key at upper 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).

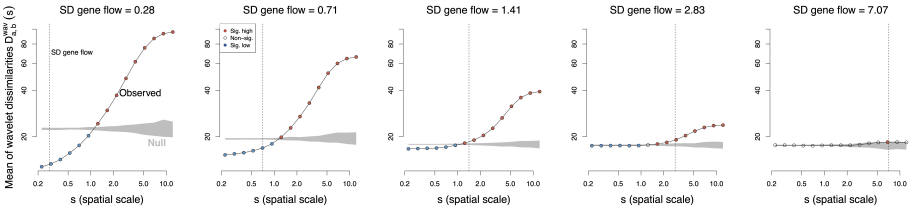

Fig. S3 Mean wavelet dissimilarities for a neutral two dimensional landscape with different levels of gene flow in simulation from each panel. 200 samples were taken from each simulation, and wavelet dissimilarities were calculated at a range of spatial scales s (connected by a solid black line) in comparison with the null expectation (gray ribbon) from permuted sample locations (2.5-97.5th percentiles of 100 permutations.

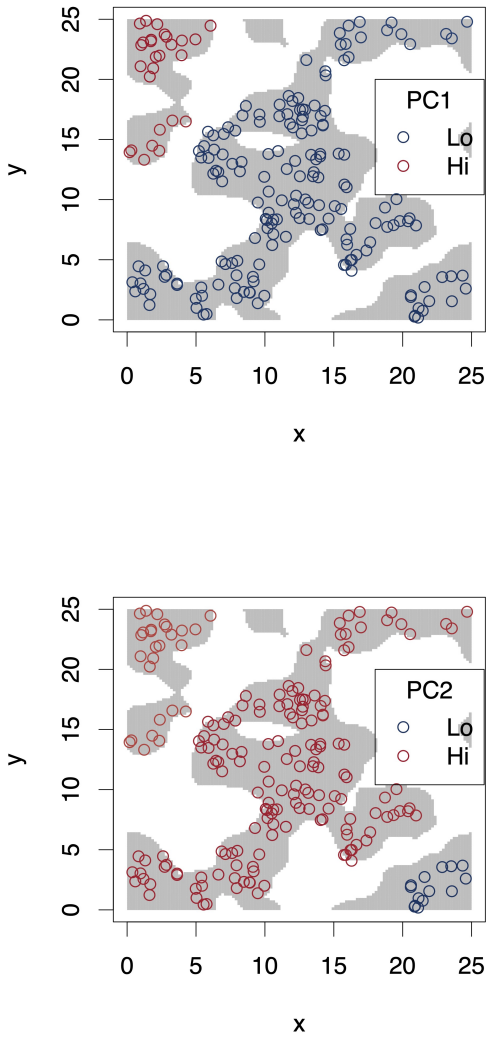

Fig. S4 Principal component analysis on 1000 random SNPs from the neutral evolution simulation on a heterogeneous landscape. Habitat is shown as gray in the background and unsuitable areas are white. Sampled individuals are circles. Colors represent the first two PCs and show how the two populations on islands in upper left and bottom right are genetically distinct.

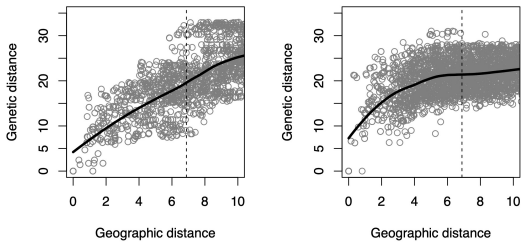

Fig. S5 Genetic distance versus geographic distance for 2 regions highlighted in Figure 3. The two regions highlighted in Figure 3B (gray boxes) are shown here, so that the left panel here corresponds to Figure 3D and the right panel to Figure 3E (the distance in PC1 space is shown in those panels of Figure 3).

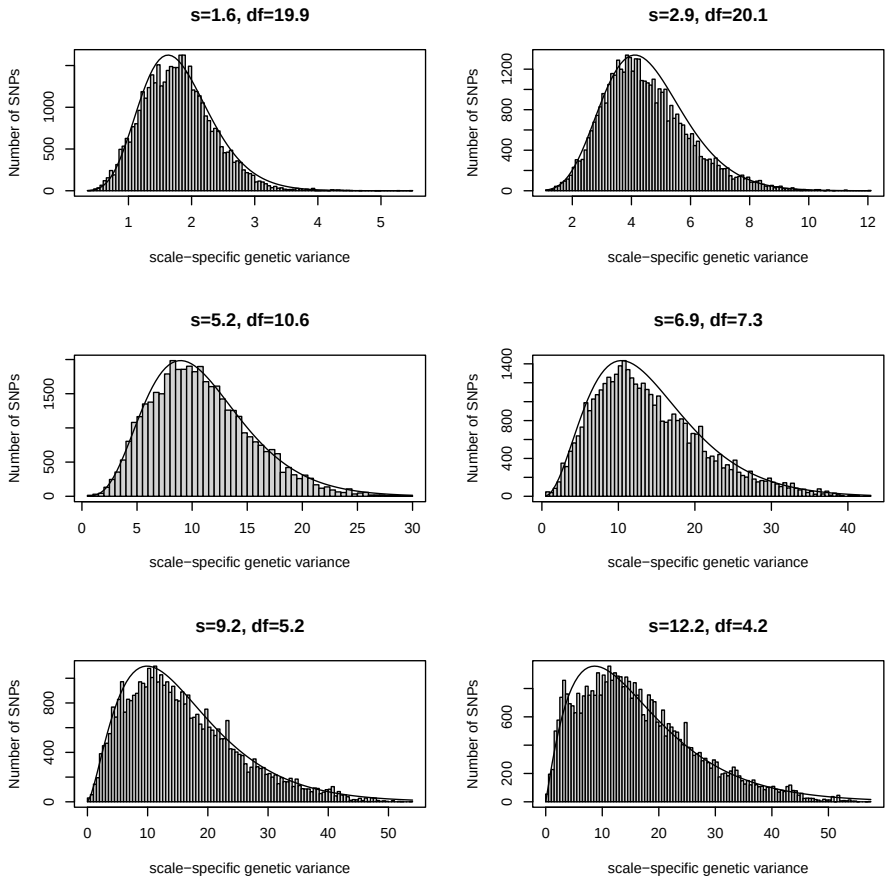

Fig. S6 Distribution of scale-specific genetic variance under a neutral scenario compared to fitted χ^2. Histogram shows SNPs and the curve shows the fitted χ^2. Here the simulation is as in Figures S2 and S3, with σ of mating and of dispersal equal to 0.02, for a SD of gene flow equal to 0.28. The degrees of freedom of the χ^2 were fit as described in the main text.

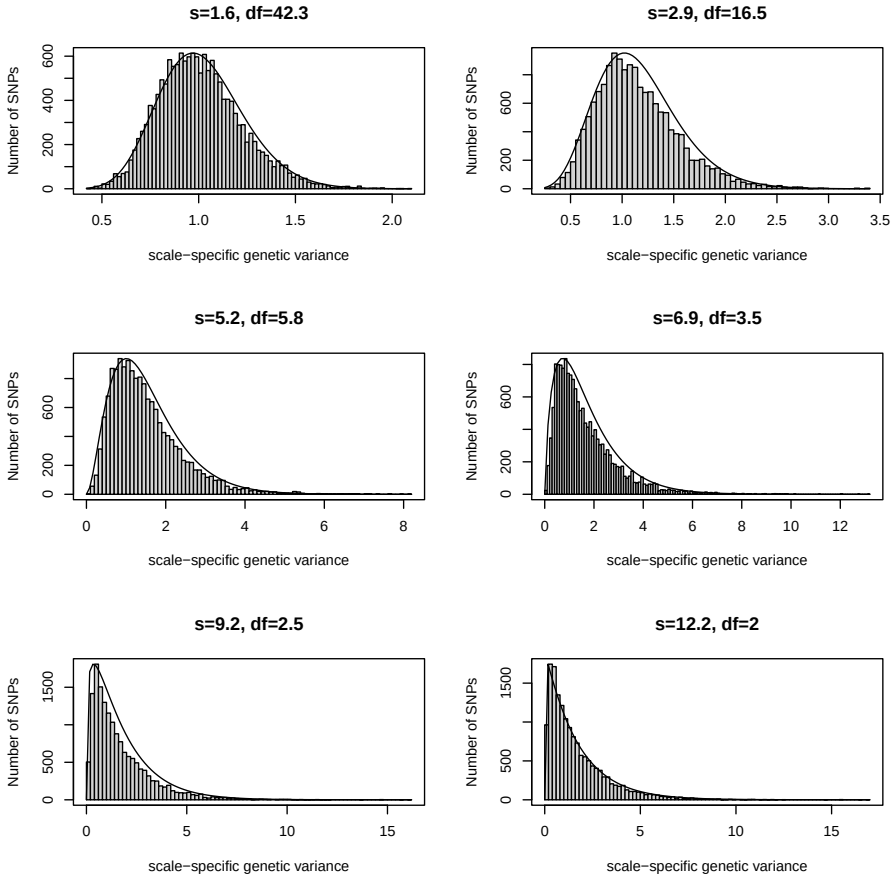

Fig. S7 Distribution of scale-specific genetic variance under a neutral scenario compared to fitted χ^2. Histogram shows SNPs and the curve shows the fitted χ^2. Here the simulation is as in Figures S2 and S3, with σ of mating and of dispersal equal to 2, for a SD of gene flow equal to 2.83. The degrees of freedom of the χ^2 were fit as described in the main text.

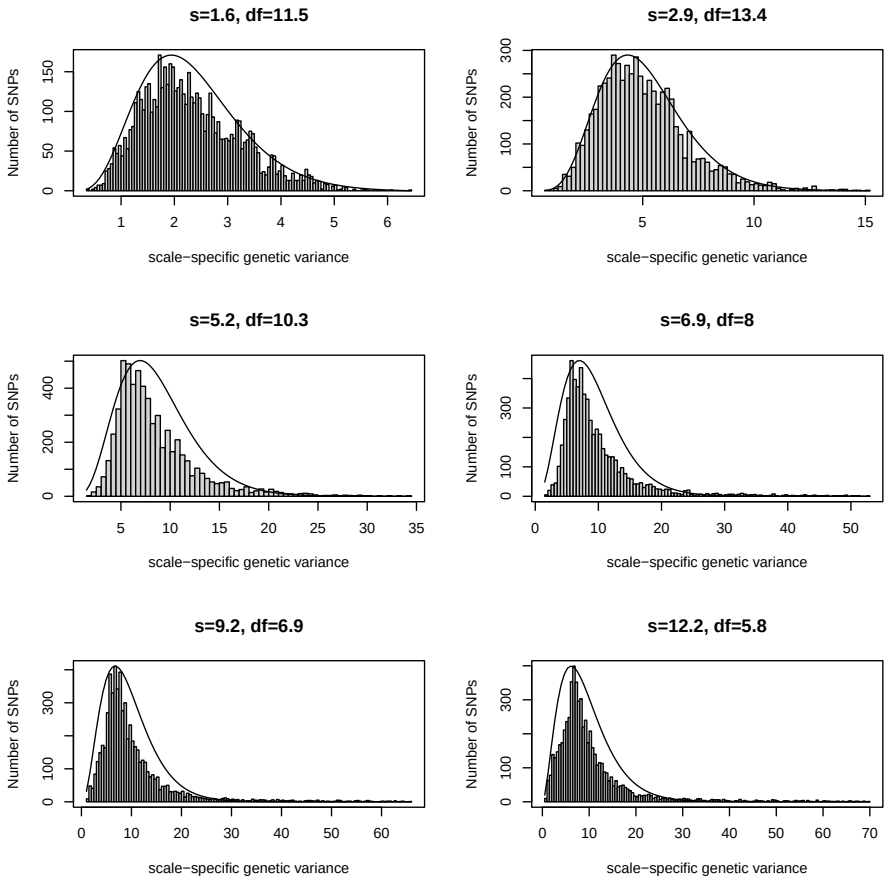

Fig. S8 Distribution of scale-specific genetic variance under a neutral scenario with recent rapid range expansion, after 3,000 years, compared to fitted χ^2. Histogram shows SNPs and the curve shows the fitted χ^2. Here the simulation is as in Figure 3. The degrees of freedom of the χ^2 were fit as described in the main text.

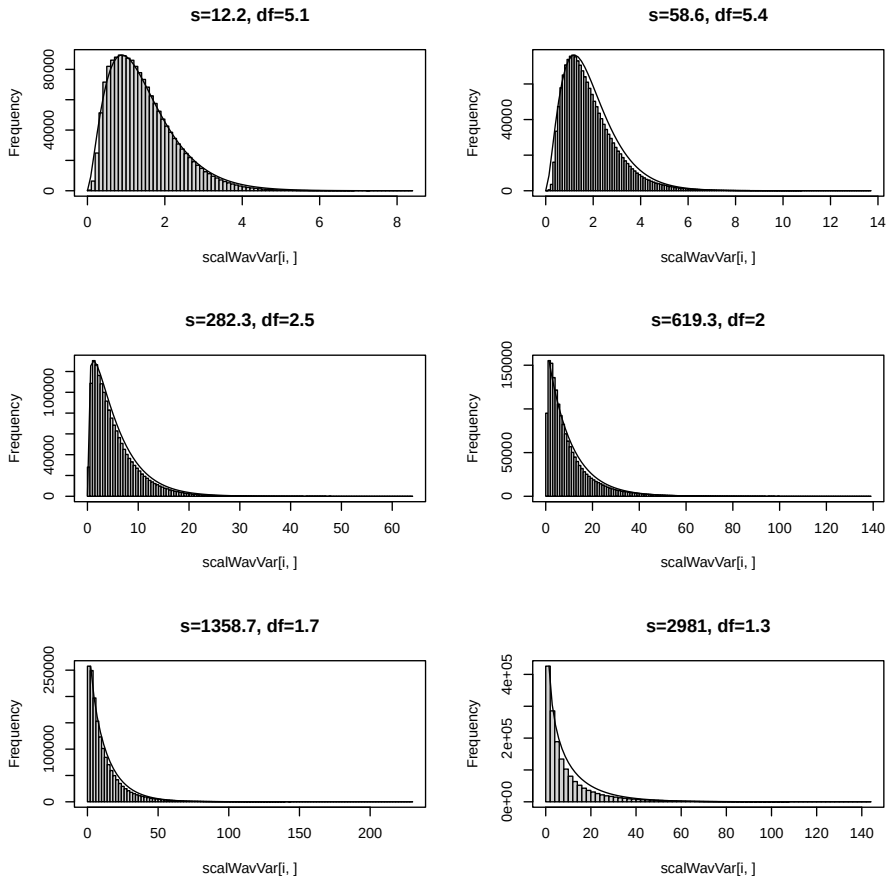

Fig. S9 Distribution of scale-specific genetic variance for Arabidopsis, compared to fitted χ^2. Histogram shows SNPs and the curve shows the fitted χ^2. The degrees of freedom of the χ^2 were fit as described in the main text.

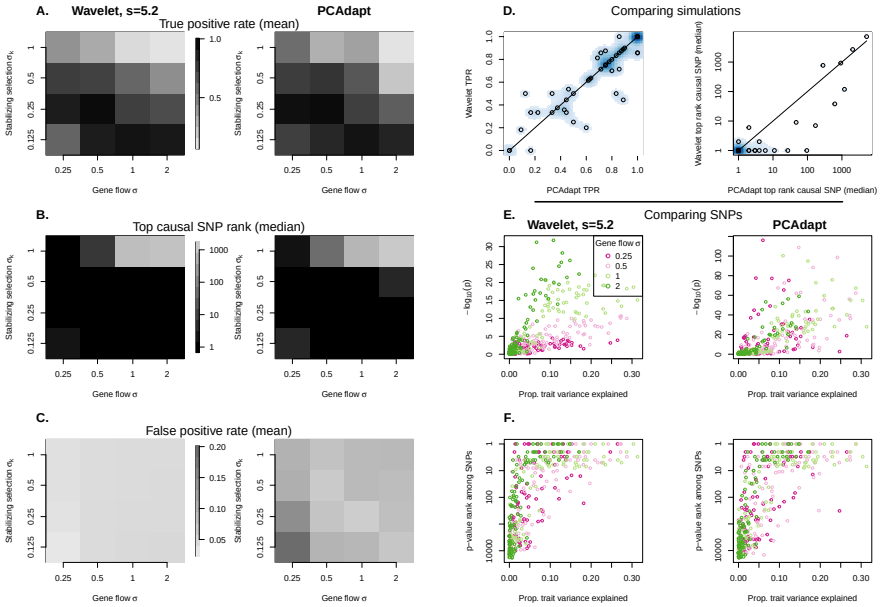

Fig. S10 Comparing the scale-specific genetic variance test with PCAdapt in simulations of adaptation to a linear environmental gradient. (A) True positive rates (nominal $p < 0.05$) for each combination of simulation parameters, the scales of mating and dispersal σ and the standard deviation of the Gaussian stabilizing selection function σ_k. (B) An alternate view of statistical power based on the median rank of the top selected SNP among all SNPs. (C) False positive rates (nominal $p < 0.05$). (D) Comparing power between the two statistical approaches for the different simulation runs. Density of points is shown in the blue scale so as to indicate where many simulations had the same result. The line indiciates a 1:1 relationship. (E-F) Individual selected SNPs in simulations, showing their nominal p values and ranks among all SNPs, colored based on σ in the simulation. The x-axis represents the proportion of total phenotypic variation among sampled individuals that was explained by the given SNP (R^2 from a linear model).

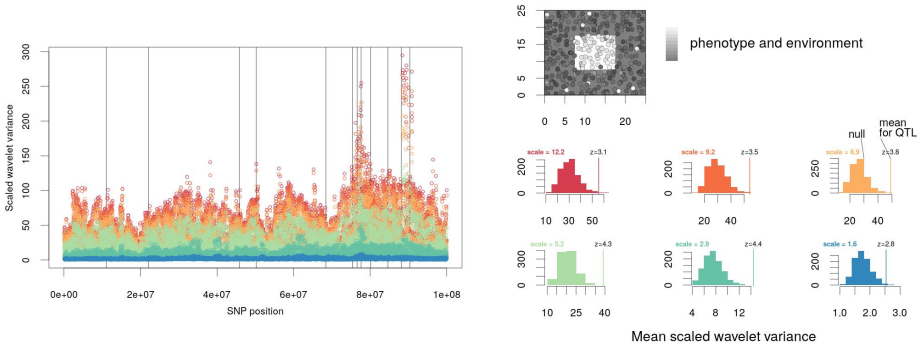

Fig. S11 Testing for selection on QTL using wavelet transforms. Comparing mean scale-specific genetic variance $var((T^{wav} f_i)(a, b, s)/sd(f_i))$ for QTL to that of random SNPs, for six different scales s (red = large scale and blue = small scale). Populations were locally adapted to a discrete habitat 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 scale-specific genetic 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 scale-specific genetic variance of QTL and its z-score.

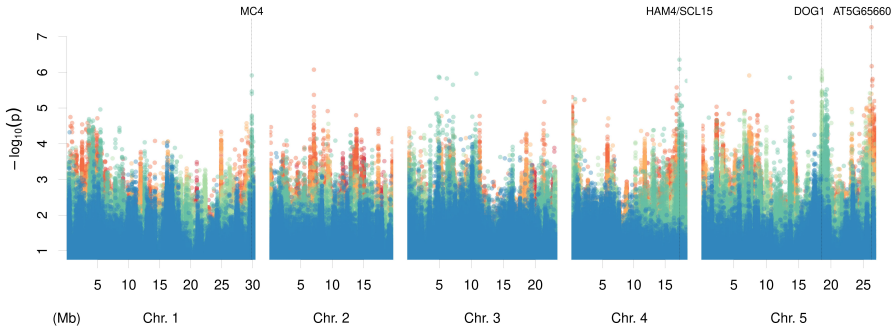

Fig. S12 Scale-specific genetic variance test results for SNPs of Arabidopsis. Scales shown go from blue (small scale) to red (large scale), specifically, ~ 12, ~ 59, ~ 282, ~ 619, ~ 1359, ~ 2980 km. Candidate genes are noted.