

¹
² Local PCA Shows How the Effect of Population Structure Differs
³ Along the Genome

⁴

⁵ Han Li¹, Peter Ralph^{1,2,3,*}

⁶ **1 Department of Molecular and Computational Biology, University of**
⁷ **Southern California, Los Angeles, CA , USA**

⁸ **2 Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA**

⁹ **3 Department of Mathematics, University of Oregon, Eugene, OR, USA**

¹⁰ * plr@uoregon.edu

¹¹ **Abstract**

¹² Population structure leads to systematic patterns in measures of mean relatedness be-
¹³ tween individuals in large genomic datasets, which are often discovered and visualized
¹⁴ using dimension reduction techniques such as principal component analysis (PCA). Mean
¹⁵ relatedness is an average of the relationships across locus-specific genealogical trees, which
¹⁶ can be strongly affected on intermediate genomic scales by linked selection and other fac-
¹⁷ tors. We show how to use local principal components analysis to describe this meso-scale
¹⁸ heterogeneity in patterns of relatedness, and apply the method to genomic data from three
¹⁹ species, finding in each that the effect of population structure can vary substantially across
²⁰ only a few megabases. In a global human dataset, localized heterogeneity is likely explained
²¹ by polymorphic chromosomal inversions. In a range-wide dataset of *Medicago truncatula*,
²² factors that produce heterogeneity are shared between chromosomes, correlate with local
²³ gene density, and may be caused by background selection or local adaptation. In a dataset
²⁴ of primarily African *Drosophila melanogaster*, large-scale heterogeneity across each chro-
²⁵ mosome arm is explained by known chromosomal inversions thought to be under recent
²⁶ selection, and after removing samples carrying inversions, remaining heterogeneity is corre-
²⁷ lated with recombination rate and gene density, again suggesting a role for linked selection.
²⁸ The visualization method provides a flexible new way to discover biological drivers of ge-
²⁹ netic variation, and its application to data highlights the strong effects that linked selection
³⁰ and chromosomal inversions can have on observed patterns of genetic variation.

31 **1 Introduction**

32 Wright [66] defined *population structure* to encompass “such matters as numbers, compo-
33 sition by age and sex, and state of subdivision”, where “subdivision” refers to restricted
34 migration between subpopulations. The phrase is also commonly used to refer to the
35 genetic patterns that result from this process, as for instance reduced mean relatedness
36 between individuals from distinct populations. However, it is not necessarily clear what
37 aspects of demography should be included in the concept. For instance, Blair [4] defines
38 *population structure* to be the sum total of “such factors as size of breeding populations,
39 periodic fluctuation of population size, sex ratio, activity range and *differential survival of*
40 *progeny*” (emphasis added). The definition is similar to Wright’s, but differs in including
41 the effects of natural selection. On closer examination, incorporating differential survival
42 or fecundity makes the concept less clear: should a randomly mating population consisting
43 of two types that are partially reproductively isolated from each other be said to show
44 population structure or not? Whatever the definition, it is clear that due to natural se-
45 lection, the effects of population structure – the *realized* patterns of genetic relatedness –
46 differ depending on which portion of the genome is being considered. For instance, strongly
47 locally adapted alleles of a gene will be selected against in migrants to different habitats,
48 increasing genetic differentiation between populations near to this gene. Similarly, newly
49 adaptive alleles spread first in local populations. These observations motivate many meth-
50 ods to search for genetic loci under selection, as for example in Huerta-Sánchez et al. [29],
51 Martin et al. [45], and Duforet-Frebbourg et al. [19].

52 These realized patterns of genetic relatedness summarize the shapes of the genealogical
53 trees at each location along the genome. Since these trees vary along the genome, so does
54 relatedness, but averaging over sufficiently many trees we hope to get a stable estimate
55 that doesn’t depend much on the genetic markers chosen. This is not guaranteed: for
56 instance, relatedness on sex chromosomes is expected to differ from the autosomes; and
57 positive or negative selection on particular loci can dramatically distort shapes of nearby
58 genealogies [3, 10, 35]. Indeed, many species show chromosome-scale variation in diversity
59 and divergence (e.g., Langley et al. [39]); species phylogenies can differ along the genome
60 due to incomplete lineage sorting, adaptive introgression and/or local adaptation (e.g.,
61 Ellegren et al. [21], Nadeau et al. [48], Pease and Hahn [54], Pool [56], Vernot and Akey
62 [63]); and theoretical expectations predict that geographic patterns of relatedness should
63 depend on selection [14].

64 Patterns in genome-wide relatedness are often summarized by applying principal com-
65 ponents analysis (PCA, Patterson et al. [53]) to the genetic covariance matrix, as pioneered
66 by Menozzi et al. [47]. The results of PCA can be related to the genealogical history of
67 the samples, such as time to most recent common ancestor and migration rate between
68 populations [46, 50], and sometimes produce “maps” of population structure that reflect
69 the samples’ geographic origin distorted by rates of gene flow [51].

70 Modeling such “background” kinship between samples is essential to genome-wide as-

71 sociation studies (GWAS, Astle and Balding [2], Price et al. [58]), and so understanding
72 variation in kinship along the genome could lead to more generally powerful methods,
73 and may be essential for doing GWAS in species with substantial heterogeneity in realized
74 patterns of mean relatedness along the genome.

75 Others have applied PCA to windows of the genome: Ma and Amos [42] used local PCA
76 much as we do to identify putative chromosomal inversions. Bryc et al. [7] and Brisbin
77 et al. [6] use PCA to infer tracts of local ancestry in recently admixed populations, but
78 by projecting each genomic window onto the axes of a single, globally-defined PCA rather
79 than doing PCA separately on each window.

80 A note on nomenclature: In this work we describe variation in patterns of relatedness
81 using local PCA, where “local” refers to proximity along the genome. A number of general
82 methods for dimensionality reduction also use a strategy of “local PCA” (e.g., Kambhatla
83 and Leen [33], Manjón et al. [44], Roweis and Saul [59], Weingessel and Hornik [65]),
84 performing PCA not on the entire dataset but instead on subsets of observations, providing
85 local pictures which are then stitched back together to give a global picture. At first
86 sight, this differs from our method in that we restrict to subsets of *variables* instead of
87 subsets of observations. However, if we flip perspectives and think of each genetic variant
88 as an observation, our method shares common threads, although our method does not
89 subsequently use adjacency along the genome, as we aim to identify similar regions that
90 may be distant.

91 It is common to describe variation along the genome of simple statistics such as F_{ST} and
92 to interpret the results in terms of the action of selection (e.g., Ellegren et al. [21], Turner
93 et al. [62]). However, a given pattern (e.g., valleys of F_{ST}) can be caused by more than
94 one biological process [8, 18], which in retrospect is unsurprising given that we are using
95 a single statistic to describe a complex process. It is also common to use methods such
96 as PCA to visualize large-scale patterns in mean genome-wide relatedness. In this paper
97 we show if and how patterns of mean relatedness vary systematically along the genome,
98 in a way particularly suited to large samples from geographically distributed populations.
99 Geographic population structure sets the stage by establishing “background” patterns of
100 relatedness; our method then describes how this structure is affected by selection and other
101 factors. Our aim is not to identify outlier loci, but rather to describe larger-scale variation
102 shared by many parts of the genome; correlation of this variation with known genomic
103 features can then be used to uncover its source.

104 2 Materials and Methods

105 As depicted in Figure 1, the general steps to the method are: (1) divide the genome into
106 windows, (2) summarize the patterns of relatedness in each window, (3) measure dissimi-
107 larity in relatedness between each pair of windows, (4) visualize the resulting dissimilarity
108 matrix using multidimensional scaling (MDS), and (5) combine similar windows to more

109 accurately visualize local effects of population structure using PCA.

110 **2.1 PCA in genomic windows**

111 To begin, we first recoded sampled genotypes as numeric matrices in the usual manner, by
112 recording the number of nonreference alleles seen at each locus for each sample. We then
113 divided the genome into contiguous segments (“windows”) and applied principal com-
114 ponent analysis (PCA) as described in McVean [46] separately to the submatrices that
115 corresponded to each window. The choice of window length entails a tradeoff between sig-
116 nal and noise, since shorter windows allow better resolution along the genome but provide
117 less precise estimates of relatedness. A method for choosing a window length to balance
118 these considerations is given in Appendix A. Precisely, denote by Z the $L \times N$ recoded
119 genotype matrix for a given window (L is the number of SNPs and N is the sample size),
120 and by \bar{Z}_s the mean of non-missing entries for allele s , so that $\bar{Z}_s = \frac{1}{n_s} \sum_j Z_{sj}$, where
121 the sum is over the n_s nonmissing genotypes. We first compute the mean-centered ma-
122 trix X , as $X_{si} = Z_{si} - \bar{Z}_s$, and preserving missingness. (This mean-centering makes the
123 result not depend on the choice of reference allele, exactly if there is no missing data,
124 and approximately otherwise.) Next, we find the covariance matrix of X , denoted C , as
125 $C_{ij} = \frac{1}{m_{ij}-1} \sum_s X_{si} X_{sj} - \frac{1}{m_{ij}(m_{ij}-1)} (\sum_s X_{si})(\sum_s X_{sj})$, where all sums are over the m_{ij}
126 sites where both sample i and sample j have nonmissing genotypes. The principal com-
127 ponents are the eigenvectors of C , normalized to have Euclidean length equal to one, and
128 ordered by magnitude of the eigenvalues.

129 The top 2–5 principal components are generally good summaries of population struc-
130 ture; for ease of visualization we usually only use the first two (referred to as PC_1 and
131 PC_2), and check that results hold using more. The above procedure can be performed on
132 any subset of the data; for future reference, denote by PC_{1j} and PC_{2j} the result after
133 applying to all SNPs in the j^{th} window. (Note, however, that our measure of dissimilarity
134 between windows does not depend on PC ordering.)

135 **2.2 Similarity of patterns of relatedness between windows**

136 We think of the local effects of population structure as being summarized by *relative*
137 position of the samples in the space defined by the top principal components. However, we
138 do not compare patterns of relatedness of different genomic regions by directly comparing
139 the PCs, since rotations or reflections of these imply identical patterns of relatedness.
140 Instead, we compare the low-dimensional approximations of the local covariance matrices
141 obtained using the top k PCs, which is invariant under ordering of the PCs , reflections, and
142 rotations and yet contains all other information about the PCs. (For results shown here,
143 we use $k = 2$; results using larger numbers of PCs were nearly identical.) Furthermore,
144 to remove the effect of artifacts such as mutation rate variation, we also rescale each
145 approximate covariance matrix to be of similar size (precisely, so that the underlying data

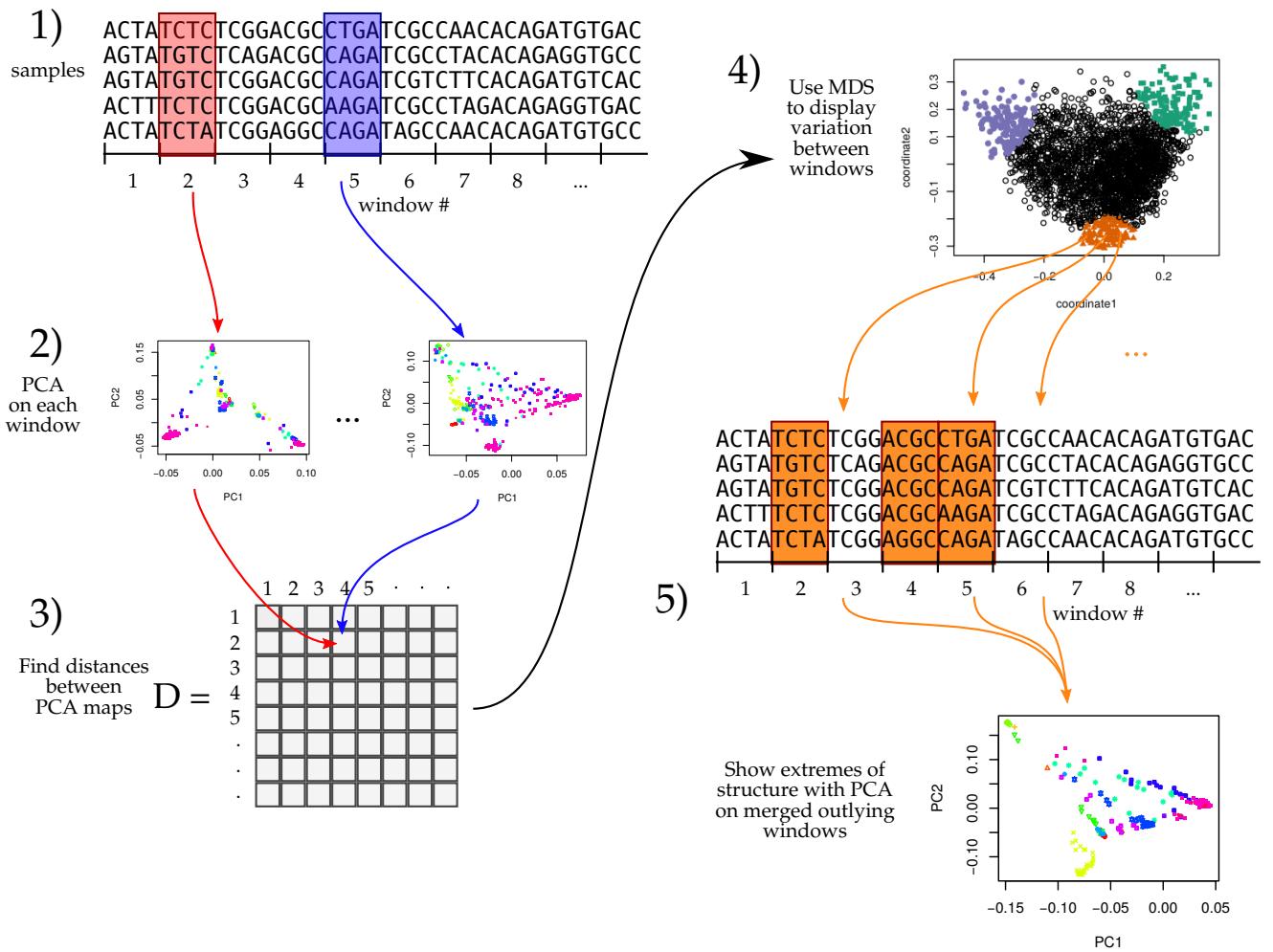


Figure 1: An illustration of the method; see Methods for details.

146 matrix has trace norm equal to one).

147 To do this, define the $N \times k$ matrix $V(i)$ so that $V(i)_{.\ell}$, the ℓ^{th} column of $V(i)$, is
148 equal to the ℓ^{th} principal component of the i^{th} window, multiplied by $(\lambda_{\ell i} / \sum_{m=1}^k \lambda_{mi})^{1/2}$,
149 where $\lambda_{\ell i}$ is the ℓ^{th} eigenvalue of the genetic covariance matrix. Then, the rescaled, rank
150 k approximate covariance matrix for the i^{th} window is

$$M(i) = \sum_{\ell=1}^k V(i)_{.\ell} V(i)_{.\ell}^T. \quad (1)$$

151 To measure the similarity of patterns of relatedness for the i^{th} window and j^{th} window,
152 we then use Euclidean distance D_{ij} between the matrices $M(i)$ and $M(j)$: $D_{ij}^2 =$
153 $\sum_{k\ell} (M(i)_{k,\ell} - M(j)_{k,\ell})^2$.

154 The goal of comparing PC plots up to rotation and reflection turned out to be equivalent
155 to comparing rank- k approximations to local covariance matrices. This suggests instead
156 directly comparing entire local covariance matrices. However, with thousands of samples
157 and tens of thousands of windows, computing the distance matrix would take months
158 of CPU time, while as defined above, D can be computed in minutes using the following
159 method. Since for square matrices A and B , $\sum_{ij} (A_{ij} - B_{ij})^2 = \sum_{ij} (A_{ij}^2 + B_{ij}^2) - 2 \text{tr}(A^T B)$,
160 then due to the orthogonality of eigenvectors and the cyclic invariance of trace, D_{ij} can be
161 computed efficiently as

$$D_{ij} = \left(\frac{\sum_{\ell=1}^k \lambda_{\ell i}^2}{(\sum_{\ell=1}^k \lambda_{\ell i})^2} + \frac{\sum_{\ell=1}^k \lambda_{\ell j}^2}{(\sum_{\ell=1}^k \lambda_{\ell j})^2} - 2 \sum_{\ell,m=1}^k (V(i)^T V(j))_{\ell m}^2 \right)^{1/2}. \quad (2)$$

162 2.3 Visualization of results

163 We use multidimensional scaling (MDS) to visualize relationships between windows as
164 summarized by the dissimilarity matrix D . MDS produces a set of m coordinates for
165 each window that give the arrangement in m -dimensional space that best recapitulates the
166 original distance matrix. For results here, we use $m = 2$ to produce one- or two-dimensional
167 visualizations of relationships between windows' patterns of relatedness.

168 We then locate variation in patterns of relatedness along the genome by choosing col-
169 lections of windows that are nearby in MDS coordinates, and map their positions along the
170 genome. A visualization of the effects of population structure across the entire collection
171 is formed by extracting the corresponding genomic regions and performing PCA on all,
172 aggregated, regions.

173 2.4 Testing

174 We tested the method using two types of simulation. First, to verify expected behavior,
175 we simulated "genomes" as an independent sequence of correlated Gaussian "genotypes",

176 using a different covariance matrix in the first quarter, middle half, and last quarter of
177 the chromosome. To verify robustness to missing data, we ran the method after randomly
178 dropping 50% of the genotypes in the first half of the genome; if the method is misled by
179 missing data, then it will distinguish the two halves of the chromosome rather than the
180 segments having different covariance matrices. Details are given in Appendix B.1.

181 To provide a realistic test, we next used forwards-time, individual-based simulations,
182 implemented using SLiM v3 [25], which are described in detail in Appendix B.2. To
183 provide realistic population structure for PCA to identify, each simulation had at least 5,000
184 diploid individuals, living across a continuous square range, with Gaussian dispersal and
185 local density-dependent competition. Each genome was modeled on human chromosome
186 7, which is 1.54×10^8 bp long, with an overall recombination rate of 1.6785 crossovers per
187 chromosome per generation. To improve speed, we used tskit [34] to record tree sequences
188 and add neutral mutations afterwards, at a rate of 10^{-9} per bp per generation. Most
189 simulations were neutral, but we also included linked selection, of two types. First, we
190 introduced selected mutations into two regions, which extended from 1/3 to 1/2 and from
191 5/6 to the end of the genome respectively. These had selection coefficients from a Gamma
192 distribution with shape 2 and mean 0.005 at a rate of 10^{-10} per bp, that were either
193 beneficial (with probability 1/30) or deleterious (otherwise). Second, to roughly model a
194 recent expansion followed by local adaptation, we introduced mutations in the same manner
195 as above, except that mutations were no longer unconditionally deleterious or beneficial:
196 each selection coefficient was multiplied by a factor depending on the spatial location of
197 the individual being evaluated, varying linearly from -1 at the left side of the range to +1
198 at the right edge. In all simulations, genome-wide PCA displayed a map of the population
199 range, as expected.

200 2.5 Datasets

201 We applied the method to genomic datasets with good geographic sampling: 380 African
202 *Drosophila melanogaster* from the Drosophila Genome Nexus [38], a worldwide dataset of
203 humans, 3,965 humans from several locations worldwide from the POPRES dataset [49],
204 and 263 *Medicago truncatula* from 24 countries around the Mediterranean basin a range-
205 wide dataset of the partially selfing weedy annual plant from the *Medicago truncatula*
206 Hapmap Project [61], as summarized in Table 1.

207 ***Drosophila melanogaster*:** We used whole-genome sequencing data from the Drosophila
208 Genome Nexus (<http://www.johnpool.net/genomes.html>, [38]), consisting of the Drosophila
209 Population Genomics Project phases 1–3 [39, 57], and additional African genomes [38]. Af-
210 ter removing 20 genomes with more than 8% missing data, we were left with 380 samples
211 from 16 countries across Africa and Europe. Since the *Drosophila* samples are from inbred
212 lines or haploid embryos, we treat the samples as haploid when recoding; regions with
213 residual heterozygosity were marked as missing in the original dataset; we also removed

214 positions with more than 20% missing data. Each chromosome arm we investigated (X,
215 2L, 2R, 3L, and 3R) has 2–3 million SNPs; PCA plots for each arm are shown in Figure
216 S3.

217 **Human:** We also used genomic data from the entire POPRES dataset [49], which has
218 array-derived genotype information for 447,267 SNPs across the 22 autosomes of 3,965
219 samples in total: 346 African-Americans, 73 Asians, 3,187 Europeans and 359 Indian
220 Asians. Since these data derive from genotyping arrays, the SNP density is much lower than
221 the other datasets, which are each derived from whole genome sequencing. We excluded
222 the sex chromosomes and the mitochondria. PCA plots for each chromosome, separately,
223 are shown in Figure S4.

224 ***Medicago truncatula*:** Finally, we used whole-genome sequencing data from the *Med-*
225 *icago truncatula* Hapmap Project [61], which has 263 samples from 24 countries, primarily
226 distributed around the Mediterranean basin. Each of the 8 chromosomes has 3–5 million
227 SNPs; PCA plots for these are shown in Figure S5. We did not use the mitochondria or
228 chloroplasts.

species	# SNPs per window	mean window length (bp)	mean # windows per chromosome	mean % variance explained by top 2 PCs
<i>Drosophila melanogaster</i>	1,000	9,019	2,674	0.53
Human	100	636,494	203	0.55
<i>Medicago truncatula</i>	10,000	102,580	467	0.50

Table 1: Descriptive statistics for each dataset used.

229 2.6 Data access

230 The methods described here are implemented in an open-source R package available at
231 https://github.com/petrelharp/local_pca, as well as scripts to perform all analyses
232 from VCF files at various parameter settings.

233 Datasets are available as follows: human (POPRES) at dbGaP with accession number
234 phs000145.v4.p2, *Medicago* at the Medicago Hapmap <http://www.medicagohapmap.org/>,
235 and *Drosophila* at the Drosophila Genome Nexus, <http://www.johnpool.net/genomes.html>.
236

237 **3 Results**

238 In all three datasets: a worldwide sample of humans, African *Drosophila melanogaster*,
239 and a rangewide sample of *Medicago truncatula*, PCA plots vary along the genome in a
240 systematic way, showing strong chromosome-scale correlations. This implies that variation
241 is due to meaningful heterogeneity in a biological process, since noise due to randomness in
242 choice of local genealogical trees is not expected to show long distance correlations. Below,
243 we discuss the results and likely underlying causes.

244 **3.1 Validation**

245 Simple non-population-based simulations with Gaussian “genotypes” showed that the method
246 performs as expected, clearly separating regions of the genome with different underlying
247 covariance matrices without being affected by extreme differences in amount of missing
248 data (Supplemental Figure S19). This simulation also verifies insensitivity to ordering of
249 top PCs, since it was performed using a covariance matrix with the top two eigenvalues
250 equal, so that the order of empirical eigenvectors (PCs) switches randomly.

251 Individual-based simulations using SLiM [25] allowed us to test the effects of recombi-
252 nation and mutation rate variation, as well as linked selection. Results are shown in
253 Supplemental Figures S16 and S17. As expected, varying recombination rate stepwise by
254 a factor of 64 did not induce patterns in the MDS visualizations correlated with recombi-
255 nation rate. Since varying mutation rate with a fixed recombination map is equivalent
256 to varying the recombination map and remapping windows, we used the same simulations
257 to show that variation in mutation rate similarly has no effect. On the other hand, a re-
258 combination map with hotspots (the HapMap human female map for chromosome 7 [31])
259 induced outliers at long regions of low recombination rate (also as expected).

260 Simulations with linked selection produced mixed results (Figure S17). The method
261 strongly identified the regions under spatially varying linked selection. It also identified
262 the regions (although less unambiguously) with constant selection and stepwise varying re-
263 combination rate, but did not clearly identify them with constant recombination rate. This
264 difference is likely because recombination rates are overall lower in the first case, leading to
265 a stronger effect of linked selection. These tests are not meant to be comprehensive survey
266 of linked selection, but only to demonstrate that linked selection can produce signals similar
267 to what we see in real data.

268 **3.2 *Drosophila melanogaster***

269 We applied the method to windows of average length 9 Kbp, across chromosome arms
270 2L, 2R, 3L, 3R and X separately. The first column of Figure 2 is a multidimensional
271 scaling (MDS) visualization of the matrix of dissimilarities between genomic windows: in
272 other words, genomic windows that are closer to each other in the MDS plot show more
273 similar patterns of relatedness. For each chromosome arm, the MDS visualization roughly

274 resembles a triangle, sometimes with additional points. Since the relative position of each
275 window in this plot shows the similarity between windows, this suggests that there are
276 at least three extreme manifestations of population structure typified by windows found
277 in the “corners” of the figure, and that other windows’ patterns of relatedness may be a
278 mixture of those extremes. The next two columns of Figure 2 respectively depict the two
279 MDS coordinates of each window, plotted against the window’s position along the genome,
280 to show how the plot of the first column is laid out along the genome. The patterns did
281 not depend on the number of PCs used (see Figure S2 for the same plot with $k = 5$ PCs),
282 and are not driven by variation in missingness (see Figure S18).

283 To help visualize how clustered windows with similar patterns of relatedness are along
284 each chromosome arm, we selected three “extreme” windows in the MDS plot and the 5%
285 of windows that are closest to it in the MDS coordinates, then highlighted these windows’
286 positions along the genome, and created PCA plots for the windows, combined. Represen-
287 tative plots are shown for three groups of windows on each chromosome arm in Figure 2
288 (groups are shown in color), and in Supplemental Figure S1 (PCA plots). The latter plots
289 are quite different, showing that genomic windows in different regions of the MDS plot
290 indeed show quite different patterns of relatedness.

291 The most striking variation in patterns of relatedness turns out to be explained by
292 several large inversions that are polymorphic in these samples, discussed in Corbett-Detig
293 and Hartl [16] and Langley et al. [39]. To depict this, Figure 3 shows the PCA plots in
294 Figure S1 recolored by the orientation of the inversion for each sample. Taking chromosome
295 arm 2L as an example, the two regions of similar, extreme patterns of relatedness shown
296 in green in the first row of Figure 2 lie directly around the breakpoints of the inversion
297 In(2L)t, and the PCA plots in the first rows of Figure 3 shows that patterns of relatedness
298 here are mostly determined by inversion orientation. The regions shown in purple on
299 chromosome 2L lie near the centromere, and have patterns of relatedness reflective of two
300 axes of variation, seen in Figures S1 and 3, which correspond roughly to latitude within
301 Africa and to degree of cosmopolitan admixture respectively (see Lack et al. [38] for more
302 about admixture in this sample). The regions shown in orange on chromosome 2L mostly
303 lie inside the inversion, and show patterns of relatedness that are a mixture between the
304 other two, as expected due to recombination within the (long) inversion [24]. Similar results
305 are found in other chromosome arms, albeit complicated by the coexistence of more than
306 one polymorphic inversion; however, each breakpoint visibly affects patterns in the MDS
307 coordinates (see vertical lines in Figure 2).

308 To see how patterns of relatedness vary in the absence of polymorphic inversions, we
309 performed the same analyses after removing, for each chromosome arm, any samples car-
310 rying inversions on that arm. In the result, shown in Supplemental Figure S6, the striking
311 peaks associated with inversion breakpoints are gone, and previously smaller-scale vari-
312 ation now dominates the MDS visualization. For instance, the majority of the variation
313 along 3L in Figure 2 is on the left end of the arm, dominated by two large peaks around the
314 inversion breakpoints; there is also a relatively small dip on the right end of the arm (near

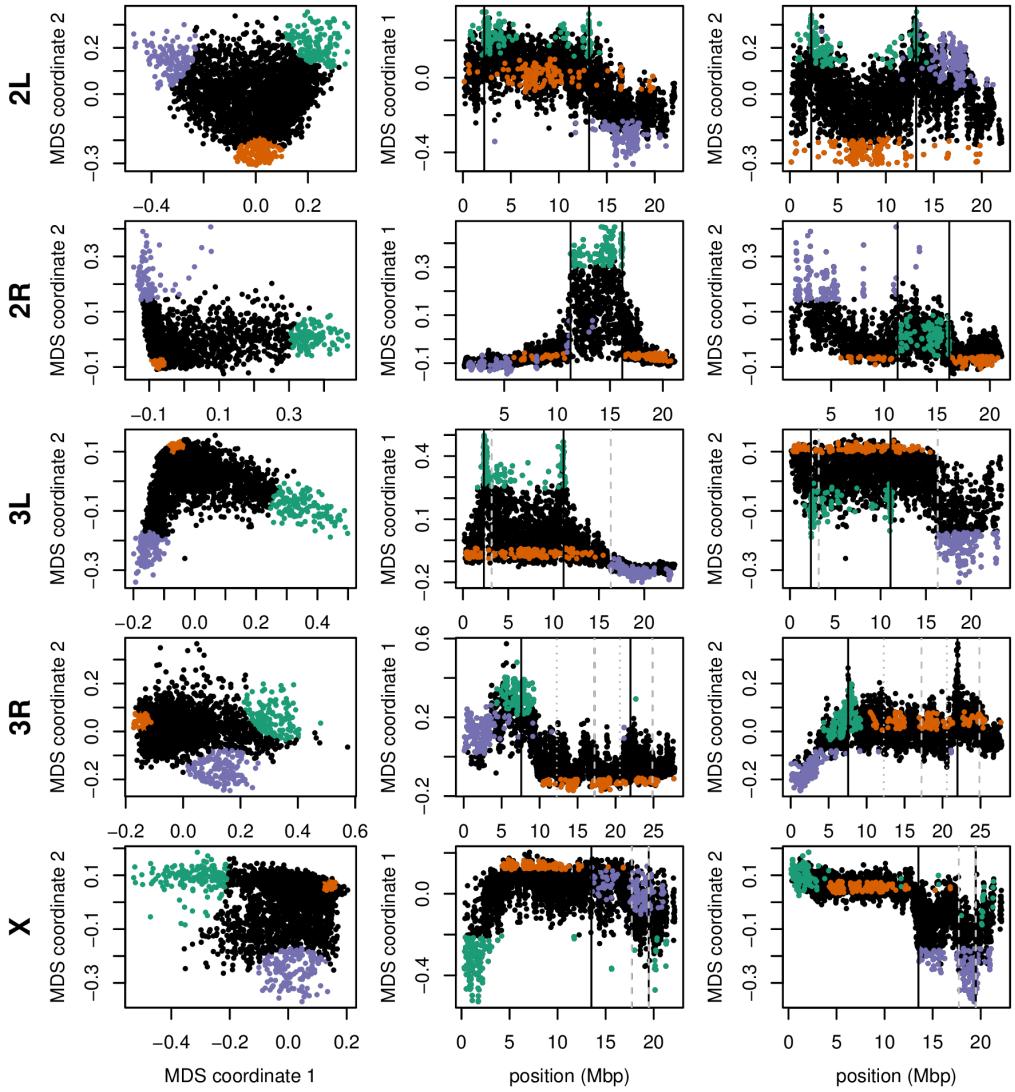


Figure 2: Variation in patterns of relatedness for windows across *Drosophila melanogaster* chromosome arms. In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the two MDS coordinates against the midpoint of each window; rows correspond to chromosome arms. Colors are consistent for plots in each row. Vertical lines show the breakpoints of known polymorphic inversions. Solid black lines are for the inversions we used in Figure 3, while dotted grey lines are for other known inversions.

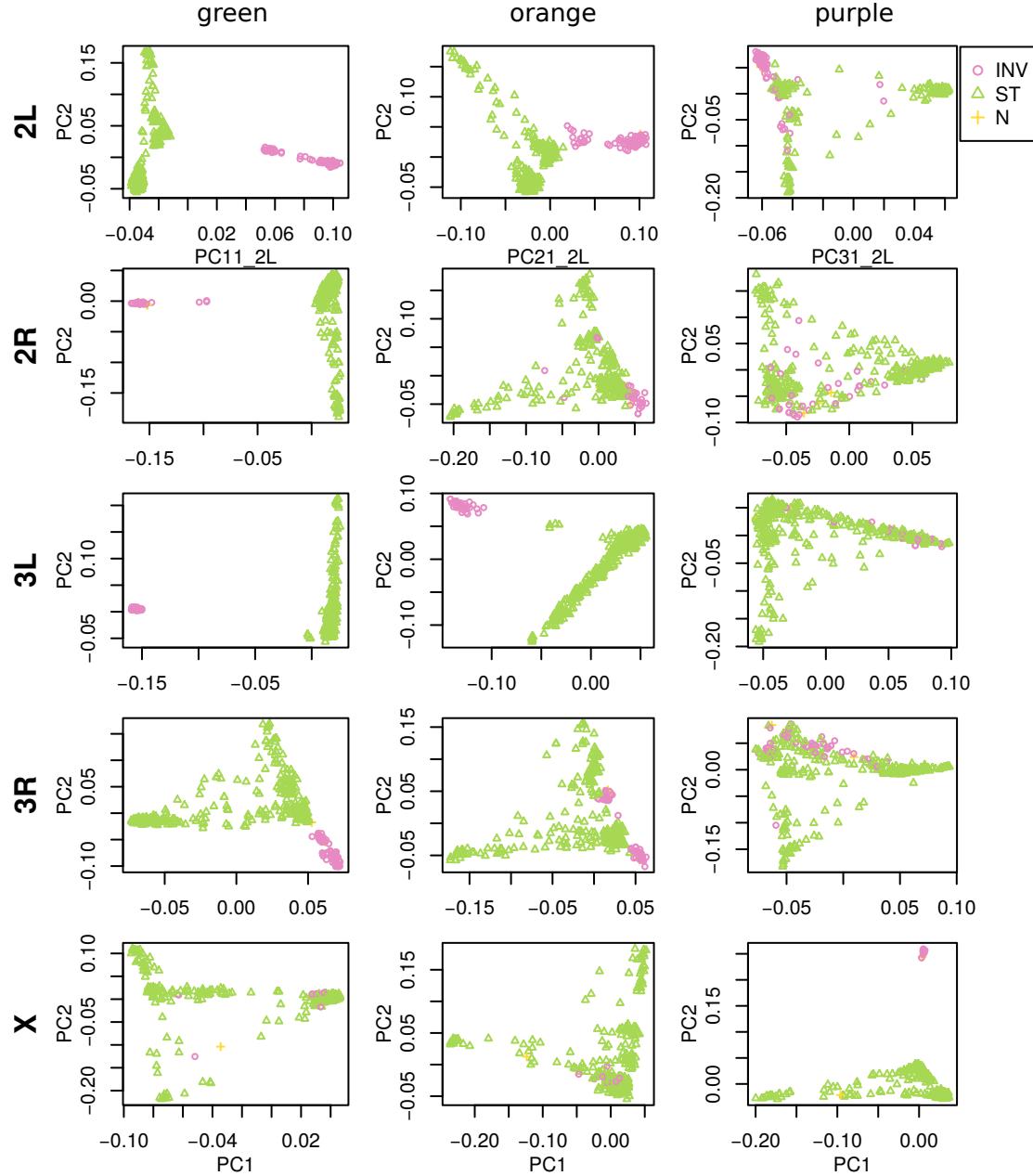


Figure 3: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange windows; and third is for purple windows. In each, samples are colored by orientation of the polymorphic inversions In(2L)t, In(2R)NS, In(3L)OK, In(3R)K and In(1)A respectively (data from [38]). In each “INV” denotes an inverted genotype, “ST” denotes the standard orientation, and “N” denotes unknown.

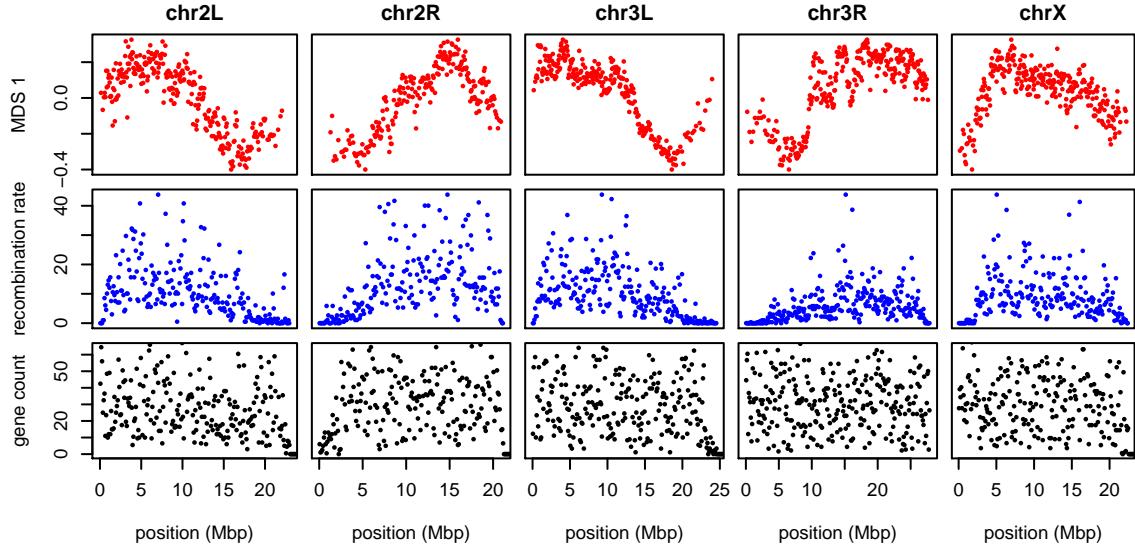


Figure 4: The effects of population structure without inversions is correlated to recombination rate in *Drosophila melanogaster*. The first plot (in red) shows the first MDS coordinate along the genome for windows of 10,000 SNPs, obtained after removing samples with inversions. (A plot analogous to Figure 2 is shown in Supplemental Figure S6.) The second plot (in blue) shows local average recombination rates in cM/Mbp, obtained as midpoint estimates for 100Kbp windows from the Drosophila recombination rate calculator [22] release 5, using rates from Comeron et al. [15]. The third plot (in black) shows the number of genes' transcription start and end sites within each 100Kbp window, divided by two. Transcription start and end sites were obtained from the RefGene table from the UCSC browser. The histone gene cluster on chromosome arm 2L is excluded.

the centromere). In contrast, Supplemental Figure S6 shows that after removing polymorphic inversions, remaining structure is dominated by the dip near the centromere. Without inversions, variation in patterns of relatedness shown in the MDS plots follows similar patterns to that previously seen in *D. melanogaster* recombination rate and diversity [39, 43]. Indeed, correlations between the recombination rate in each window and the position on the first MDS coordinate are highly significant (Spearman's $\rho = 0.54$, $p < 2 \times 10^{-16}$; Figures 4 and S7). This is consistent with the hypothesis that variation is due to selection, since the strength of linked selection increases with local gene density, measured in units of recombination distance. The number of genes – measured as the number of transcription start and end sites within each window – was not significantly correlated with MDS coordinate ($p = 0.22$).

326 **3.3 Human**

327 As we did for the *Drosophila* data, we applied our method separately to all 22 human
328 autosomes. On each, variation in patterns of relatedness was dominated by a small number
329 of windows having similar patterns of relatedness to each other that differed dramatically
330 from the rest of the chromosome. These may be primarily inversions: outlying windows
331 coincide with three of the six large polymorphic inversions described in Antonacci et al. [1],
332 notably a particularly large, polymorphic inversion on 8p23 (Figure 5). Similar plots for
333 all chromosomes are shown in Supplementary Figures S8, S9, and S10. PCA plots of many
334 outlying windows show a characteristic trimodal shape (shown for chromosome 8 in Figure
335 S11), presumably distinguishing samples having each of the three diploid genotypes for each
336 inversion orientation (although we do not have data on orientation status). This trimodal
337 shape has been proposed as a method to identify inversions [42], but distinguishing this
338 hypothesis from others, such as regions of low recombination rate, would require additional
339 data.

340 We also applied the method on all 22 autosomes together, and found that, remarkably,
341 the inversion on chromosome 8 is still the most striking outlying signal (Figure S12).
342 Further investigation with a denser set of SNPs, allowing a finer genomic resolution, may
343 yield other patterns.

344 **3.4 *Medicago truncatula***

345 Unlike the other two species, the method applied separately on all eight chromosomes of
346 *Medicago truncatula* showed similar patterns of gradual change in patterns of relatedness
347 across each chromosome, with no indications of chromosome-specific patterns. This con-
348 sistency suggests that the factor affecting the population structure for each chromosome
349 is the same, as might be caused by varying strengths of linked selection. To verify that
350 variation in the effects of population structure is shared across chromosomes, we applied
351 the method to all chromosomes together. Results for chromosome 3 are shown in Figure
352 6, and other chromosomes are similar: across chromosomes, the high values of the first
353 MDS coordinate coincide with the position of the heterochromatic regions surrounding
354 the centromere, which often have lower gene density and may therefore be less subject to
355 linked selection. To verify that this is a possible explanation, we counted the number of
356 genes found in each window using gene models in Mt4.0 from jcvi.org [61], which are
357 shown juxtaposed with the first MDS coordinate of each window in Figure 7, and are sig-
358 nificantly correlated, as shown in Supplemental Figure S13. (Values shown are the number
359 of start and end positions of each predicted mRNA transcript, divided by two, assigned
360 to the nearest window.) However, other genomic features, such as distance to centromere
361 show roughly the same patterns, so we cannot rule out alternative hypotheses. In particu-
362 lar, fine-scale recombination rate estimates are not available in a form mappable to Mt4.0
363 coordinates (although those in Paape et al. [52] appear visually similar).

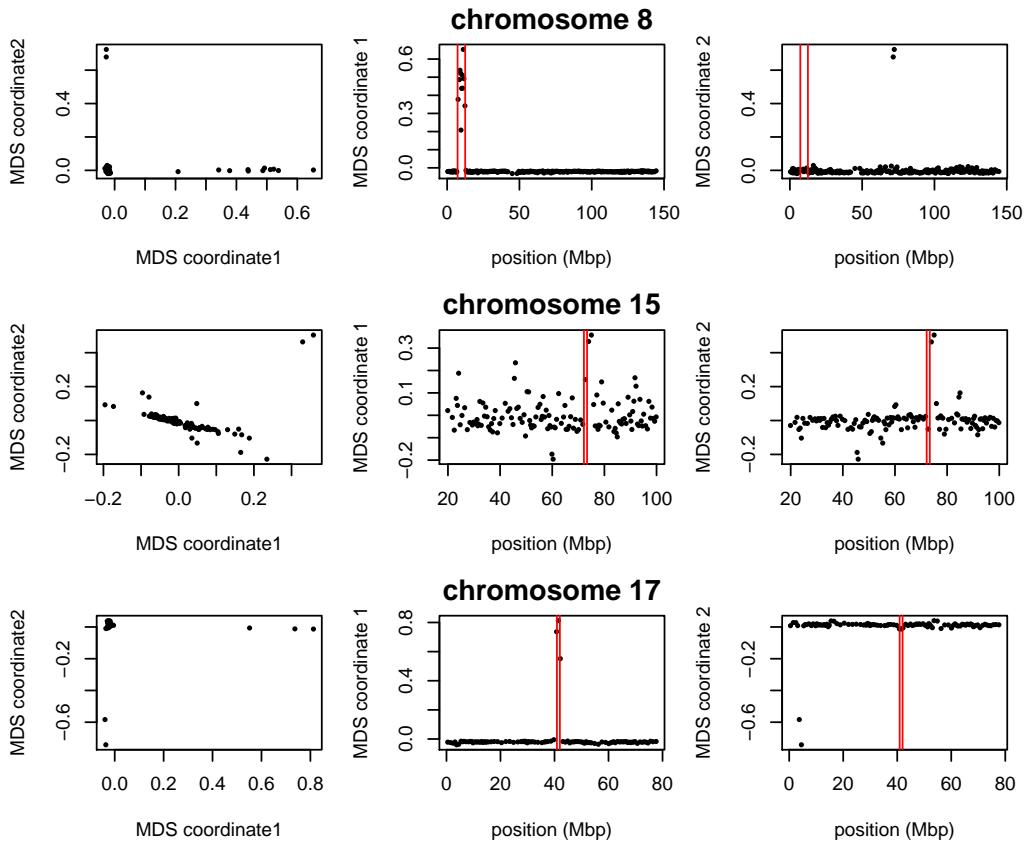


Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in each plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position (midpoint) along the chromosome. Rows, from top to bottom show chromosomes 8, 15, and 17. The vertical red lines show the breakpoints of known inversions from Antonacci et al. [1].

364 The results were highly consistent across window sizes, window types (SNPs or bp),
365 and number of PCs (Table S2).

366 **4 Discussion**

367 Our investigations have found substantial variation in the patterns of relatedness formed
368 by population structure across the genomes of three diverse species, revealing distinct bi-
369 ological processes driving this variation in each species. More investigation, particularly
370 on more species and datasets, will help to uncover what aspects of species history can ex-
371 plain these differences. With growing appreciation of the heterogeneous effects of selection
372 across the genome, especially the importance of adaptive introgression and hybrid specia-
373 tion [5, 23, 30, 56, 60], local adaptation [40, 64], and inversion polymorphisms [36, 37], local
374 PCA may prove to be a useful exploratory tool to discover important genomic features.

375 We now discuss possible implications of this variation in the effects of population struc-
376 ture, the impact of various parameter choices in implementing the method, and possible
377 additional applications.

378 **Chromosomal inversions** A major driver of variation in patterns of relatedness in two
379 datasets we examined are inversions. This may be common, but the example of *Medicago*
380 *truncatula* shows that polymorphic inversions are not ubiquitous. PCA has been proposed
381 as a method for discovering inversions [42]; however, the signal left by inversions likely
382 cannot be distinguished from long haplotypes under balancing selection or simply regions
383 of reduced recombination without additional lines of evidence. Inversions show up in our
384 method because across the inverted region, most gene trees share a common split that
385 dates back to the origin of the inversion. However, in many applications, inversions are a
386 nuisance. For instance, SMARTPCA [53] reduces their effect on PCA plots by regressing
387 out the effect of linked SNPs on each other. Removing samples with the less common
388 orientation of each inversion reduced, but did not eliminate, the signal of inversions seen in
389 the *Drosophila melanogaster* dataset, demonstrating that the genomic effects of transiently
390 polymorphic inversions may outlast the inversions themselves.

391 **The effect of selection** It seems that the variation in patterns of relatedness we see in
392 the *Medicago truncatula* and *Drosophila melanogaster* datasets must be explained some-
393 how by linked selection – the effects of neutral processes are not expected to show such
394 large, chromosome-scale correlations. Furthermore, the selection must be affecting many
395 targets across the genome, since we see similar effects across long distances (even distinct
396 chromosomes). For this reason, the most likely candidate may be selection against linked
397 deleterious mutations, known as “background selection” [10, 13]. Informally, background
398 selection reduces the number of potential contributors to the gene pool in regions of the
399 genome with many possible deleterious mutations [28]; for this reason, if it acts in a spatial

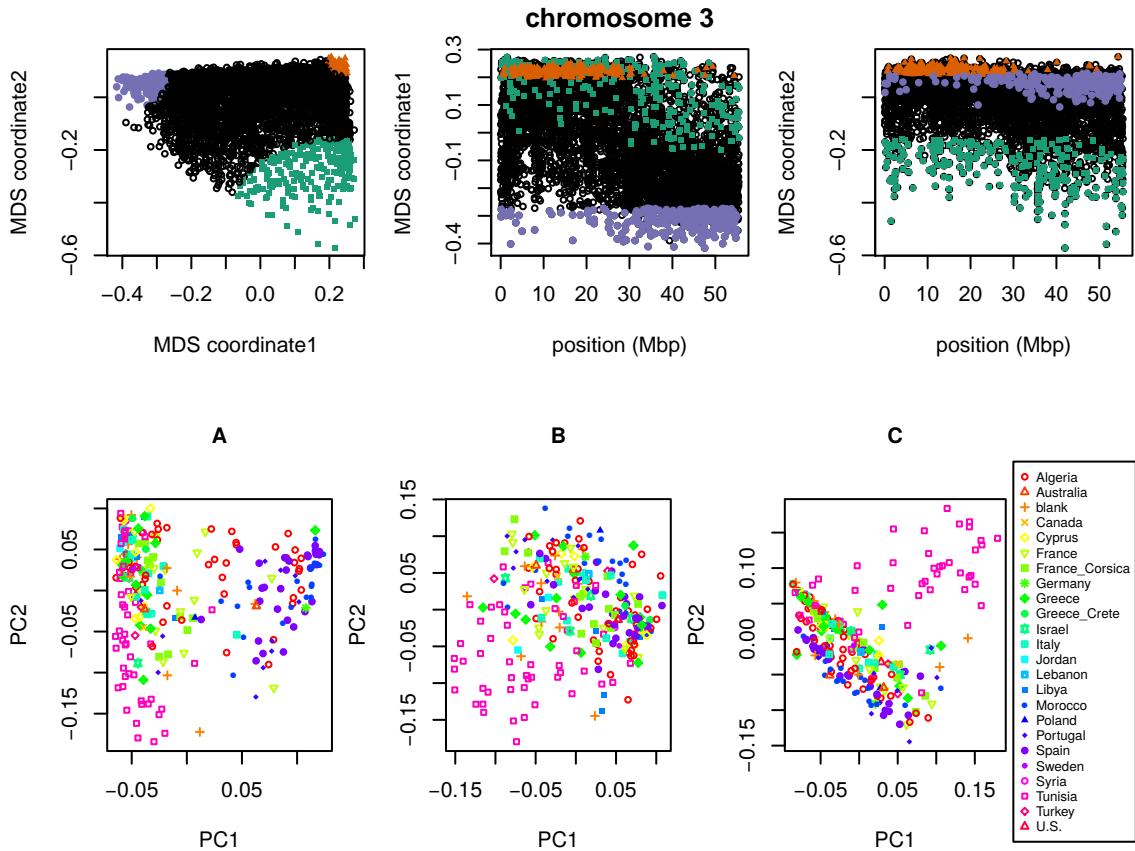


Figure 6: **(top)** MDS visualization of patterns of relatedness on *M. truncatula* chromosome 3, with corresponding PCA plots. Each point in the plot represents a window; the structure revealed by the MDS plot is strongly clustered along the chromosome, with windows in the upper-right corner of the MDS plot (colored red) clustered around the centromere, windows in the upper-left corner (purple) furthest from the centromere, and the remaining corner (green) intermediate. Plots for remaining chromosomes are shown in Supplemental Figure S14. **(bottom)** PCA plots for the sets of genomic windows colored (A) green, (B) orange, and (C) purple in the top figure. Each point corresponds to a sample, colored by country of origin. Plots for remaining chromosomes are shown in Supplemental Figure S15.

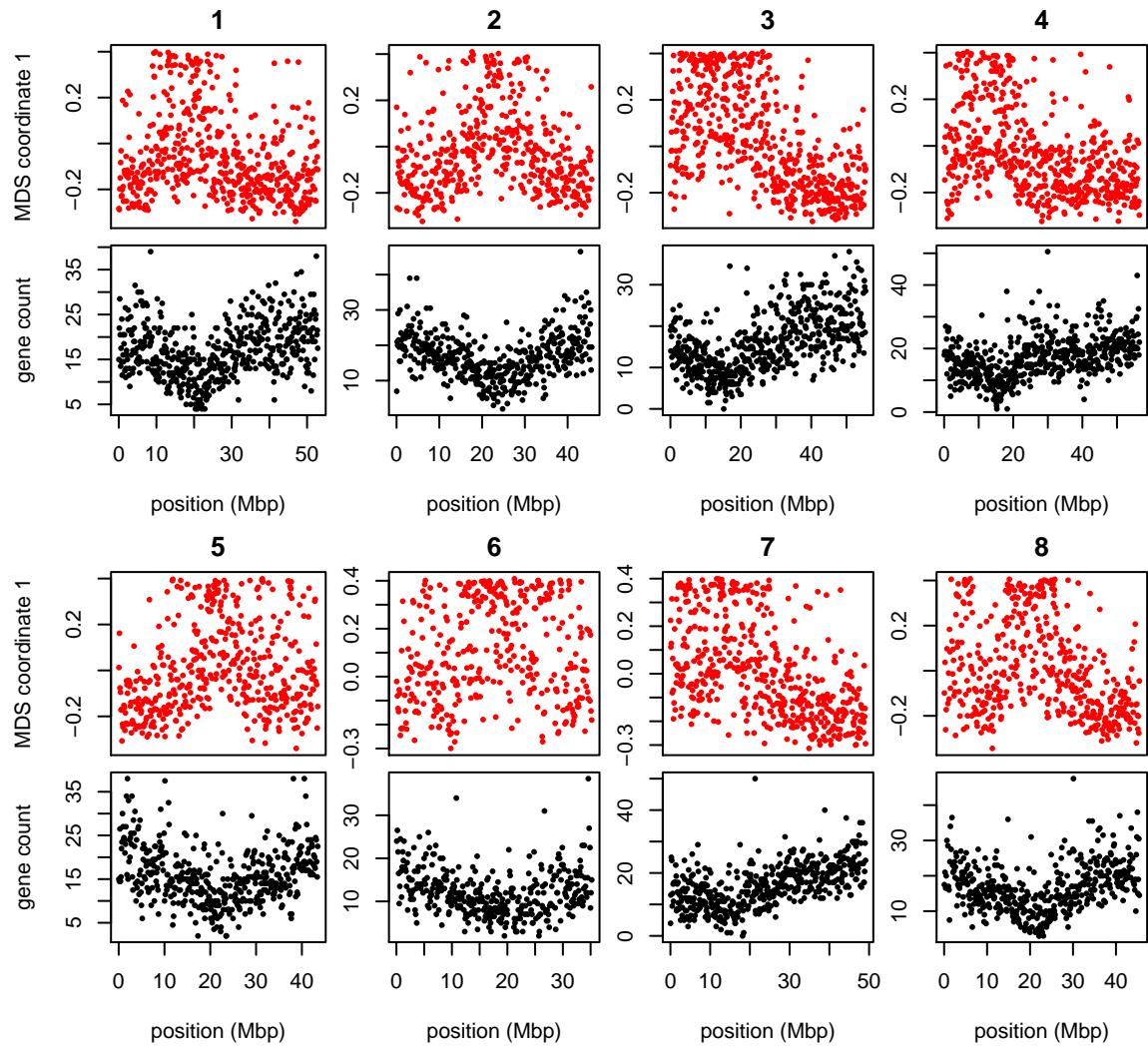


Figure 7: MDS coordinate and gene density for each window in the *Medicago* genome, for chromosomes 1–8 (numbered above each pair of figures). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene count for each window against the position of each window.

400 context, it is expected to induce samples from nearby locations to cluster together more
401 frequently. Therefore, regions of the genome harboring many targets of local adaptation
402 may show similar patterns, since migrant alleles in these regions will be selected against,
403 and so locally gene trees will more closely reflect spatial proximity.

404 A related possibility is that variation in patterns of relatedness is due to recent ad-
405 mixture between previously separated populations, the effects of which were not uniform
406 across the genome due to selection. For instance, it has been hypothesized that large-scale
407 variation in amount of introgressed Neanderthal DNA along the genome is due to selec-
408 tion against Neanderthal genes, leading to greater introgression in regions of lower gene
409 density [26, 32]. African *Drosophila melanogaster* are known to have a substantial amount
410 of recently introgressed genome from “cosmopolitan” sources; if selection regularly favors
411 genes from one origin, this could lead to substantial variation in patterns of relatedness
412 correlated with local gene density.

413 There has been substantial debate over the relative impacts of different forms of se-
414 lection [8, 11, 12, 17, 26, 27, 45, 54, 55]. These have been difficult to disentangle in part
415 because for the most part theory makes predictions which are only strictly valid in ran-
416 domly mating (i.e., unstructured) populations, and it is unclear to what extent the spatial
417 structure observed in most real populations will affect these predictions. It may be possible
418 to design more powerful statistics that make stronger use of spatial information.

419 **Parameter choices** There are several choices in the method that may in principle affect
420 the results. As with whole-genome PCA, the choice of samples is important, as variation
421 not strongly represented in the sample will not be discovered. The effects of strongly
422 imbalanced sampling schemes are often corrected by dropping samples in overrepresented
423 groups; but downweighting may be a better option that does not discard data (and here we
424 present a method to do this). Next, the choice of window size may be important, although
425 in our applications results were not sensitive to this, indicating that we can see variation
426 on a sufficiently fine scale. Finally, which collections of genomic regions are compared to
427 each other (steps 3 and 4 in Figure 1), along with the method used to discover common
428 structure, will affect results. We used MDS, applied to either each chromosome separately
429 or to the entire genome; for instance, human inversions are clearly visible as outliers when
430 compared to the rest of their chromosome, but genome-wide, their signal is obscured by
431 the numerous other signals of comparable strength.

432 Besides window length, there is also the question of how to choose windows. In these
433 applications we have used nonoverlapping windows with equal numbers of polymorphic
434 sites. Alternatively, windows could be chosen to have equal length in genetic distance, so
435 that each would have roughly the same number of independent trees. However, we found
436 little change in results when using different window sizes or when measuring windows in
437 physical distance (in bp).

438 Finally, our software allows different choices for how many PCs to use in approximating
439 structure of each window (k in equation 1), and how many MDS coordinates to use when

440 describing the distance matrix between windows, but in our exploration, changing these has
441 not produced dramatically different results. These are all part of more general techniques
442 in dimension reduction and high-dimensional data visualization; we encourage the user to
443 experiment.

444 **Applications** So-called cryptic relatedness between samples has been one of the major
445 sources of confounding in genome-wide association studies (GWAS) and so methods must
446 account for it by modeling population structure or kinship [2, 67]. Modern “mixed model”
447 methods [e.g. 41] account for this with either a single, genome-wide kinship matrix or
448 one constructed using only sites unlinked to the focal SNP. Since the effects of population
449 structure is not constant along the genome, this could in principle lead to an inflation of
450 false positives in parts of the genome with stronger population structure than the genome-
451 wide average. A method such as ours might be used to estimate local kinship matrices,
452 thus providing a more sensitive correction, although doing so without removing the signal
453 itself could be challenging. Fortunately, in our human dataset this does not seem likely
454 to have a strong effect: most variation is due to small, independent regions, possibly
455 primarily inversions, and so may not have a major effect on GWAS. In the other species we
456 examined, particularly *Drosophila melanogaster*, treating population structure as a single
457 quantity would entail a substantial loss of power, and could potentially be misleading.

458 Acknowledgements

459 We are indebted to John Pool, Russ Corbett-Detig, Matilde Cordeiro, and Peter Chang
460 for assistance with obtaining data and interpreting results (especially inversion status of
461 *D. melanogaster* samples). Jaime Ashander and Jerome Kelleher provided assistance in
462 performing the simulations. Thanks also go to Yaniv Brandvain, Barbara Engelhardt,
463 Charles Langley, Graham Coop, and Jeremy Berg for helpful comments and for encouraging
464 the project.

465 Disclosure declaration

466 The authors declare no conflicts of interest.

467 References

- 468 [1] Francesca Antonacci, Jeffrey M Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia
469 Siswara, Zhaoshi Jiang, and Evan E Eichler. Characterization of six human disease-
470 associated inversion polymorphisms. *Human molecular genetics*, 18(14):2555–2566,
471 2009.

- 472 [2] William Astle and David J. Balding. Population structure and cryptic relatedness in
473 genetic association studies. *Statistical Science*, 24(4):451–471, 11 2009. doi: 10.1214/
474 09-STS307. URL <http://dx.doi.org/10.1214/09-STS307>.
- 475 [3] Nicholas H. Barton. Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, 355
476 (1403):1553–1562, November 2000. doi: 10.1098/rstb.2000.0716. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1692896/>.
- 477 [4] Albert P. Blair. Population structure in toads. *The American Naturalist*, 77(773):563–
478 568, 1943. ISSN 00030147, 15375323. URL <http://www.jstor.org/stable/2457848>.
- 479 [5] Yaniv Brandvain, Amanda M. Kenney, Lex Flagel, Graham Coop, and Andrea L.
480 Sweigart. Speciation and introgression between *Mimulus nasutus* and *Mimulus gutta-*
481 *tus*. *PLoS Genet*, 10(6):e1004410, 06 2014. doi: 10.1371/journal.pgen.1004410. URL
482 <http://dx.doi.org/10.1371%2Fjournal.pgen.1004410>.
- 483 [6] A. Brisbin, K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. Degenhardt, A. Reynolds,
484 H. Ostrer, J. G. Mezey, and C. D. Bustamante. PCAdmix: principal components-
485 based assignment of ancestry along each chromosome in individuals with admixed
486 ancestry from two or more populations. *Hum. Biol.*, 84(4):343–364, August 2012.
- 487 [7] K Bryc, A Auton, M R Nelson, J R Oksenberg, S L Hauser, S Williams, A Froment,
488 J M Bodo, C Wambebe, S A Tishkoff, and C D Bustamante. Genome-wide patterns
489 of population structure and admixture in West Africans and African Americans. *Proc
490 Natl Acad Sci U S A*, 107(2):786–791, January 2010. doi: 10.1073/pnas.0909559107.
491 URL <http://www.ncbi.nlm.nih.gov/pubmed/20080753>.
- 492 [8] R Burri, A Nater, T Kawakami, C F Mugal, P I Olason, L Smeds, A Suh, L Du-
493 toit, S Bureš, L Z Garamszegi, S Hogner, J Moreno, A Qvarnström, M Ružić, S A
494 Sæther, G P Sætre, J Török, and H Ellegren. Linked selection and recombination rate
495 variation drive the evolution of the genomic landscape of differentiation across the spe-
496 ciation continuum of Ficedula flycatchers. *Genome Res*, 25(11):1656–1665, November
497 2015. doi: 10.1101/gr.196485.115. URL <https://www.ncbi.nlm.nih.gov/pubmed/26355005>.
- 498 [9] Frank M.T.A. Busing, Erik Meijer, and Rien Van Der Leeden. Delete-m jackknife
499 for unequal m. *Statistics and Computing*, 9(1):3–8, 1999. ISSN 0960-3174. doi:
500 10.1023/A:1008800423698. URL <http://dx.doi.org/10.1023/A%3A1008800423698>.
- 501 [10] B Charlesworth, M T Morgan, and D Charlesworth. The effect of deleterious mutations
502 on neutral molecular variation. *Genetics*, 134(4):1289–1303, August 1993. URL <http://www.genetics.org/content/134/4/1289>.

- 506 [11] B Charlesworth, M Nordborg, and D Charlesworth. The effects of local selection,
507 balanced polymorphism and background selection on equilibrium patterns of genetic
508 diversity in subdivided populations. *Genet Res*, 70(2):155–174, October 1997. URL
509 <https://www.ncbi.nlm.nih.gov/pubmed/9449192>.
- 510 [12] Brian Charlesworth. The effects of deleterious mutations on evolution at linked sites.
511 *Genetics*, 190(1):5–22, January 2012. doi: 10.1534/genetics.111.134288. URL <http://www.ncbi.nlm.nih.gov/pubmed/22219506>.
- 513 [13] Brian Charlesworth. Background selection 20 years on: The Wilhelmine E. Key 2012
514 invitational lecture. *Journal of Heredity*, 104(2):161–171, 2013. doi: 10.1093/jhered/
515 ess136. URL <http://jhered.oxfordjournals.org/content/104/2/161.abstract>.
- 516 [14] Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton. The effects of
517 genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution,
518 and Systematics*, 34(1):99–125, 2003. doi: 10.1146/annurev.ecolsys.34.
519 011802.132359. URL <http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.34.011802.132359>.
- 521 [15] J M Comeron, R Ratnappan, and S Bailin. The many landscapes of recombination
522 in *Drosophila melanogaster*. *PLoS Genet*, 8(10), 2012. doi: 10.1371/journal.pgen.
523 1002905. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3469467/>.
- 524 [16] Russell B Corbett-Detig and Daniel L Hartl. Population genomics of inversion poly-
525 morphisms in *Drosophila melanogaster*. *PLoS Genet*, 8(12):e1003056, 2012.
- 526 [17] Russell B. Corbett-Detig, Daniel L. Hartl, and Timothy B. Sackton. Natural selection
527 constrains neutral diversity across a wide range of species. *PLoS Biol*, 13(4):e1002112,
528 04 2015. doi: 10.1371/journal.pbio.1002112. URL <http://dx.doi.org/10.1371%2Fjournal.pbio.1002112>.
- 530 [18] T E Cruickshank and M W Hahn. Reanalysis suggests that genomic islands of specia-
531 tion are due to reduced diversity, not reduced gene flow. *Mol Ecol*, 23(13):3133–3157,
532 07 2014. doi: 10.1111/mec.12796. URL <https://www.ncbi.nlm.nih.gov/pubmed/24845075>.
- 534 [19] Nicolas Duforet-Frebourg, Keurcien Luu, Guillaume Laval, Eric Bazin, and
535 Michael G.B. Blum. Detecting genomic signatures of natural selection with principal
536 component analysis: Application to the 1000 genomes data. *Molecular Biology and
537 Evolution*, 2015. doi: 10.1093/molbev/msv334. URL <http://mbe.oxfordjournals.org/content/early/2016/01/12/molbev.msv334.abstract>.
- 539 [20] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for
540 Industrial and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319. URL
541 <http://pubs.siam.org/doi/abs/10.1137/1.9781611970319>.

- 542 [21] Hans Ellegren, Linnea Smeds, Reto Burri, Pall I. Olason, Niclas Backstrom, Takeshi
543 Kawakami, Axel Kunstner, Hannu Makinen, Krystyna Nadachowska-Brzyska, Anna
544 Qvarnstrom, Severin Uebbing, and Jochen B. W. Wolf. The genomic landscape of
545 species divergence in Ficedula flycatchers. *Nature*, 491(7426):756–760, November
546 2012. ISSN 00280836. doi: 10.1038/nature11584. URL <http://dx.doi.org/10.1038/nature11584>.
- 548 [22] Anna-Sophie Fiston-Lavier, Nadia D. Singh, Mikhail Lipatov, and Dmitri A. Petrov.
549 *Drosophila melanogaster* recombination rate calculator. *Gene*, 463(1–2):18 – 20, 2010.
550 ISSN 0378-1119. doi: <http://dx.doi.org/10.1016/j.gene.2010.04.015>. URL <http://www.sciencedirect.com/science/article/pii/S0378111910001769>.
- 552 [23] B M Fitzpatrick, J R Johnson, D K Kump, J J Smith, S R Voss, and H B Shaffer.
553 Rapid spread of invasive genes into a threatened native species. *Proc Natl Acad Sci
554 U S A*, 107(8):3606–3610, February 2010. doi: 10.1073/pnas.0911802107. URL <http://www.ncbi.nlm.nih.gov/pubmed/20133596>.
- 556 [24] Rafael F. Guerrero, François Rousset, and Mark Kirkpatrick. Coalescent patterns
557 for chromosomal inversions in divergent populations. *Philosophical Transactions
558 of the Royal Society B: Biological Sciences*, 367(1587):430–438, 2011. ISSN 0962-
559 8436. doi: 10.1098/rstb.2011.0246. URL <http://rstb.royalsocietypublishing.org/content/367/1587/430>.
- 561 [25] Benjamin C. Haller and Philipp W. Messer. SLiM 2: Flexible, interactive forward ge-
562 netic simulations. *Molecular Biology and Evolution*, 34(1):230–240, 2017. doi: 10.1093/
563 molbev/msw211. URL [/brokenurl#+http://dx.doi.org/10.1093/molbev/msw211](http://dx.doi.org/10.1093/molbev/msw211).
- 564 [26] Kelley Harris and Rasmus Nielsen. The genetic cost of Neanderthal introgression. *Ge-
565 netics*, 203(2):881–891, June 2016. URL <http://www.genetics.org/content/203/2/881>.
- 567 [27] P W Hedrick. Adaptive introgression in animals: examples and comparison to new
568 mutation and standing variation as sources of adaptive variation. *Mol Ecol*, 22(18):
569 4606–4618, September 2013. doi: 10.1111/mec.12415. URL <https://www.ncbi.nlm.nih.gov/pubmed/23906376>.
- 571 [28] R R Hudson and N L Kaplan. Deleterious background selection with recombi-
572 nation. *Genetics*, 141(4):1605–1617, December 1995. URL <http://www.genetics.org/content/141/4/1605>.
- 574 [29] Emilia Huerta-Sánchez, Michael DeGiorgio, Luca Pagani, Ayele Tarekegn, Rose-
575 mary Ekong, Tiago Antao, Alexia Cardona, Hugh E. Montgomery, Gianpiero L.
576 Cavalleri, Peter A. Robbins, Michael E. Weale, Neil Bradman, Endashaw Bekele,

- 577 Toomas Kivisild, Chris Tyler-Smith, and Rasmus Nielsen. Genetic signatures re-
578 veal high-altitude adaptation in a set of Ethiopian populations. *Molecular Biol-*
579 *ogy and Evolution*, 30(8):1877–1888, 2013. doi: 10.1093/molbev/mst089. URL
580 <http://mbe.oxfordjournals.org/content/30/8/1877.abstract>.
- 581 [30] Matthew B. Hufford, Pesach Lubinksy, Tanja Pyhäjärvi, Michael T. Devengenzo, Nor-
582 man C. Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild intro-
583 gression in maize. *PLoS Genet*, 9(5):e1003477, 05 2013. doi: 10.1371/journal.pgen.
584 1003477. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1003477>.
- 585 [31] International HapMap Consortium, K A Frazer, D G Ballinger, D R Cox, D A Hinds,
586 L L Stuve, R A Gibbs, J W Belmont, A Boudreau, P Hardenbol, S M Leal, S Paster-
587 nak, D A Wheeler, T D Willis, F Yu, H Yang, C Zeng, Y Gao, H Hu, W Hu, C Li,
588 W Lin, S Liu, H Pan, X Tang, J Wang, W Wang, J Yu, B Zhang, Q Zhang, H Zhao,
589 H Zhao, J Zhou, S B Gabriel, R Barry, B Blumenstiel, A Camargo, M Defelice,
590 M Faggart, M Goyette, S Gupta, J Moore, H Nguyen, R C Onofrio, M Parkin, J Roy,
591 E Stahl, E Winchester, L Ziaugra, D Altshuler, Y Shen, Z Yao, W Huang, X Chu,
592 Y He, L Jin, Y Liu, Y Shen, W Sun, H Wang, Y Wang, Y Wang, X Xiong, L Xu,
593 M M Waye, S K Tsui, H Xue, J T Wong, L M Galver, J B Fan, K Gunderson, S S
594 Murray, A R Oliphant, M S Chee, A Montpetit, F Chagnon, V Ferretti, M Leboeuf,
595 J F Olivier, M S Phillips, S Roumy, C Sallée, A Verner, T J Hudson, P Y Kwok,
596 D Cai, D C Koboldt, R D Miller, L Pawlikowska, P Taillon-Miller, M Xiao, L C Tsui,
597 W Mak, Y Q Song, P K Tam, Y Nakamura, T Kawaguchi, T Kitamoto, T Mori-
598 zono, A Nagashima, Y Ohnishi, A Sekine, T Tanaka, T Tsunoda, P Deloukas, C P
599 Bird, M Delgado, E T Dermitzakis, R Gwilliam, S Hunt, J Morrison, D Powell, B E
600 Stranger, P Whittaker, D R Bentley, M J Daly, P I de Bakker, J Barrett, Y R Chretien,
601 J Maller, S McCarroll, N Patterson, I Pe'er, A Price, S Purcell, D J Richter, P Sabeti,
602 R Saxena, S F Schaffner, P C Sham, P Varilly, D Altshuler, L D Stein, L Krishnan,
603 A V Smith, M K Tello-Ruiz, G A Thorisson, A Chakravarti, P E Chen, D J Cutler,
604 C S Kashuk, S Lin, G R Abecasis, W Guan, Y Li, H M Munro, Z S Qin, D J Thomas,
605 G McVean, A Auton, L Bottolo, N Cardin, S Eyheramendy, C Freeman, J Marchini,
606 S Myers, C Spencer, M Stephens, P Donnelly, L R Cardon, G Clarke, D M Evans, A P
607 Morris, B S Weir, T Tsunoda, J C Mullikin, S T Sherry, M Feolo, A Skol, H Zhang,
608 C Zeng, H Zhao, I Matsuda, Y Fukushima, D R Macer, E Suda, C N Rotimi, C A
609 Adebamowo, I Ajayi, T Aniagwu, P A Marshall, C Nkwodimma, C D Royal, M F
610 Leppert, M Dixon, A Peiffer, R Qiu, A Kent, K Kato, N Niikawa, I F Adewole, B M
611 Knoppers, M W Foster, E W Clayton, J Watkin, R A Gibbs, J W Belmont, D Muzny,
612 L Nazareth, E Sodergren, G M Weinstock, D A Wheeler, I Yakub, S B Gabriel, R C
613 Onofrio, D J Richter, L Ziaugra, B W Birren, M J Daly, D Altshuler, R K Wilson, L L
614 Fulton, J Rogers, J Burton, N P Carter, C M Clee, M Griffiths, M C Jones, K McLay,
615 R W Plumb, M T Ross, S K Sims, D L Willey, Z Chen, H Han, L Kang, M Godbout,
616 J C Wallenburg, P L'Archevêque, G Bellemare, K Saeki, H Wang, D An, H Fu, Q Li,

- 617 Z Wang, R Wang, A L Holden, L D Brooks, J E McEwen, M S Guyer, V O Wang,
618 J L Peterson, M Shi, J Spiegel, L M Sung, L F Zacharia, F S Collins, K Kennedy,
619 R Jamieson, and J Stewart. A second generation human haplotype map of over 3.1
620 million SNPs. *Nature*, 449(7164):851–861, October 2007. doi: 10.1038/nature06258.
621 URL <http://www.ncbi.nlm.nih.gov/pubmed/17943122>.
- 622 [32] Ivan Juric, Simon Aeschbacher, and Graham Coop. The strength of selection
623 against Neanderthal introgression. *bioRxiv*, 2016. doi: 10.1101/030148. URL
624 <http://biorkxiv.org/content/early/2016/07/22/030148>.
- 625 [33] N. Kambhatla and T. K. Leen. Dimension reduction by local principal compo-
626 nent analysis. *Neural Computation*, 9(7):1493–1516, July 1997. ISSN 0899-7667.
627 doi: 10.1162/neco.1997.9.7.1493. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6795533.
- 629 [34] Jerome Kelleher, Kevin Thornton, Jaime Ashander, and Peter Ralph. Efficient pedi-
630 gree recording for fast population genetics simulation. *bioRxiv*, 2018. doi: 10.1101/
631 248500. URL <https://www.biorkxiv.org/content/early/2018/06/07/248500>.
- 632 [35] Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking
633 along a recombining chromosome. *Genetics*, 160(2):765–777, 2002. URL <http://www.genetics.org/cgi/content/abstract/160/2/765>.
- 635 [36] Mark Kirkpatrick. How and why chromosome inversions evolve. *PLoS Biol*, 8(9), 2010.
636 doi: 10.1371/journal.pbio.1000501. URL <http://www.ncbi.nlm.nih.gov/pubmed/20927412>.
- 638 [37] Mark Kirkpatrick and Brian Barrett. Chromosome inversions, adaptive cassettes and
639 the evolution of species' ranges. *Molecular Ecology*, 2015. ISSN 1365-294X. doi:
640 10.1111/mec.13074. URL <http://dx.doi.org/10.1111/mec.13074>.
- 641 [38] Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russ-
642 ell B Corbett-Detig, Kristian A Stevens, Charles H Langley, and John E Pool.
643 The Drosophila genome nexus: a population genomic resource of 623 *Drosophila*
644 *melanogaster* genomes, including 197 from a single ancestral range population. *Ge-
645 netics*, 199(4):1229–1241, 2015.
- 646 [39] C H Langley, K Stevens, C Cardeno, Y C Lee, D R Schrider, J E Pool, S A Langley,
647 C Suarez, R B Corbett-Detig, B Kolaczkowski, S Fang, P M Nista, A K Holloway,
648 A D Kern, C N Dewey, Y S Song, M W Hahn, and D J Begun. Genomic variation
649 in natural populations of *Drosophila melanogaster*. *Genetics*, 192(2):533–598, Octo-
650 ber 2012. doi: 10.1534/genetics.112.142018. URL <http://www.ncbi.nlm.nih.gov/pubmed/22673804>.

- 652 [40] Thomas Lenormand. Gene flow and the limits to natural selection. *Trends in*
653 *Ecology & Evolution*, 17(4):183 – 189, 2002. ISSN 0169-5347. doi: DOI:10.1016/
654 S0169-5347(02)02497-7. URL <http://www.sciencedirect.com/science/article/pii/S0169534702024977>.
- 655 [41] P R Loh, G Tucker, B K Bulik-Sullivan, B J Vilhjálmsson, H K Finucane, R M
656 Salem, D I Chasman, P M Ridker, B M Neale, B Berger, N Patterson, and A L
657 Price. Efficient Bayesian mixed-model analysis increases association power in large
658 cohorts. *Nat Genet*, 47(3):284–290, March 2015. doi: 10.1038/ng.3190. URL <https://www.ncbi.nlm.nih.gov/pubmed/25642633>.
- 661 [42] J Ma and C I Amos. Investigation of inversion polymorphisms in the human genome
662 using principal components analysis. *PLoS One*, 7(7), 2012. doi: 10.1371/journal.
663 pone.0040224. URL <http://www.ncbi.nlm.nih.gov/pubmed/22808122>.
- 664 [43] Trudy F. C. Mackay, Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F.
665 Ayroles, Dianhui Zhu, Sonia Casillas, Yi Han, Michael M. Magwire, Julie M. Cridland,
666 Mark F. Richardson, Robert R. H. Anholt, Maite Barron, Crystal Bess, Kerstin Petra
667 Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan,
668 Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N. Jhangiani, Katherine
669 W. Jordan, Fremiet Lara, Faye Lawrence, Sandra L. Lee, Pablo Librado, Raquel S.
670 Linheiro, Richard F. Lyman, Aaron J. Mackey, Mala Munidasa, Donna Marie Muzny,
671 Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel
672 Ramia, Jeffrey G. Reid, Stephanie M. Rollmann, Julio Rozas, Nehad Saada, Lavanya
673 Turlapati, Kim C. Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu,
674 Casey M. Bergman, Kevin R. Thornton, David Mittelman, and Richard A. Gibbs. The
675 *Drosophila melanogaster* genetic reference panel. *Nature*, 482(7384):173–178, February
676 2012. ISSN 00280836. URL <http://dx.doi.org/10.1038/nature10811>.
- 677 [44] José V Manjón, Pierrick Coupé, Luis Concha, Antonio Buades, D Louis Collins, and
678 Montserrat Robles. Diffusion weighted image denoising using overcomplete local PCA.
679 *PloS one*, 8(9):e73021, 2013.
- 680 [45] Simon Henry Martin, Markus Moest, Wiliam J Palmer, Camilo Salazar, W. Owen
681 McMillan, Francis M Jiggins, and Chris D Jiggins. Natural selection and genetic
682 diversity in the butterfly *Heliconius melpomene*. *Genetics*, 203(1):525–541, May 2016.
683 doi: 10.1101/042796. URL <http://www.genetics.org/content/203/1/525>.
- 684 [46] Gil McVean. A genealogical interpretation of principal components analysis. *PLoS*
685 *Genet*, 5(10):e1000686, 2009.
- 686 [47] P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies
687 in Europeans. *Science*, 201(4358):786–792, September 1978. URL <http://www.ncbi.nlm.nih.gov/pubmed/356262>.

- 689 [48] N J Nadeau, A Whibley, R T Jones, J W Davey, K K Dasmahapatra, S W Baxter,
690 M A Quail, M Joron, R H ffrench Constant, M L Blaxter, J Mallet, and C D Jiggins.
691 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-
692 scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci*, 367(1587):343–353,
693 February 2012. doi: 10.1098/rstb.2011.0198. URL <http://www.ncbi.nlm.nih.gov/pubmed/22201164>.
- 695 [49] M R Nelson, K Bryc, K S King, A Indap, A R Boyko, J Novembre, L P Bri-
696 ley, Y Maruyama, D M Waterworth, G Waerber, P Vollenweider, J R Oksenberg,
697 S L Hauser, H A Stirnadel, J S Kooner, J C Chambers, B Jones, V Mooser, C D
698 Bustamante, A D Roses, D K Burns, M G Ehm, and E H Lai. The Population
700 Reference Sample, POPRES: a resource for population, disease, and pharmacolog-
701 ical genetics research. *Am J Hum Genet*, 83(3):347–358, September 2008. doi:
702 10.1016/j.ajhg.2008.08.005. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556436/?tool=pubmed>.
- 703 [50] John Novembre and Matthew Stephens. Interpreting principal component analyses of
704 spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.
- 705 [51] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko,
706 Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al.
707 Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.
- 708 [52] Timothy Paape, Peng Zhou, Antoine Branca, Roman Briskine, Nevin Young, and
709 Peter Tiffin. Fine-scale population recombination rates, hotspots, and correlates of
710 recombination in the *Medicago truncatula* genome. *Genome Biology and Evolution*,
711 4(5):726–737, 2012. doi: 10.1093/gbe/evs046. URL <http://gbe.oxfordjournals.org/content/4/5/726.abstract>.
- 713 [53] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanal-
714 sis. *PLoS Genetics*, 2(12):e190, 12 2006. doi: 10.1371/journal.pgen.0020190. URL
715 <http://dx.plos.org/10.1371%2Fjournal.pgen.0020190>.
- 716 [54] J B Pease and M W Hahn. More accurate phylogenies inferred from low-recombination
717 regions in the presence of incomplete lineage sorting. *Evolution*, 67(8):2376–2384,
718 August 2013. doi: 10.1111/evo.12118. URL <http://www.ncbi.nlm.nih.gov/pubmed/23888858>.
- 720 [55] Tanya N. Phung, Christian D. Huber, and Kirk E. Lohmueller. Determining the effect
721 of natural selection on linked neutral divergence across species. *PLOS Genetics*, 12
722 (8):1–27, 08 2016. doi: 10.1371/journal.pgen.1006199. URL <https://doi.org/10.1371/journal.pgen.1006199>.

- 724 [56] John E Pool. The mosaic ancestry of the Drosophila Genetic Reference Panel and the
725 *D. melanogaster* reference genome reveals a network of epistatic fitness interactions.
726 *Molecular Biology and Evolution*, 32(12):3236–3251, 2015. doi: 10.1101/014837. URL
727 <http://mbe.oxfordjournals.org/content/32/12/3236.abstract>.
- 728 [57] John E. Pool, Russell B. Corbett-Detig, Ryuichi P. Sugino, Kristian A. Stevens,
729 Charis M. Cardeno, Marc W. Crepeau, Pablo Duchen, J. J. Emerson, Perot Sae-
730 lao, David J. Begun, and Charles H. Langley. Population genomics of sub-Saharan
731 *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet*,
732 8(12):1–24, 12 2012. doi: 10.1371/journal.pgen.1003080. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1003080>.
- 734 [58] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A
735 Shadick, and David Reich. Principal components analysis corrects for stratification in
736 genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.
- 737 [59] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally
738 linear embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075. doi: 10.1126/
739 science.290.5500.2323. URL <http://science.sciencemag.org/content/290/5500/2323>.
- 741 [60] Fabian Staubach, Anna Lorenc, Philipp W. Messer, Kun Tang, Dmitri A. Petrov,
742 and Diethard Tautz. Genome patterns of selection and introgression of haplotypes in
743 natural populations of the house mouse (*Mus musculus*). *PLoS Genet*, 8(8):e1002891,
744 08 2012. doi: 10.1371/journal.pgen.1002891. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1002891>.
- 746 [61] Haibao Tang, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan,
747 Shiguo Zhou, Laurent Gentzbittel, Kevin L Childs, Mark Yandell, Heidrun Gundlach,
748 et al. An improved genome release (version mt4. 0) for the model legume *Medicago*
749 *truncatula*. *BMC genomics*, 15(1):1, 2014.
- 750 [62] T L Turner, M W Hahn, and S V Nuzhdin. Genomic islands of speciation in *Anopheles*
751 *gambiae*. *PLoS Biol*, 3(9), September 2005. doi: 10.1371/journal.pbio.0030285. URL
752 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1260762/>.
- 753 [63] Benjamin Vernot and Joshua M. Akey. Resurrecting surviving neandertal lineages from
754 modern human genomes. *Science*, 2014. doi: 10.1126/science.1245938. URL <http://www.sciencemag.org/content/early/2014/01/28/science.1245938.abstract>.
- 756 [64] Ian J. Wang and Gideon S. Bradburd. Isolation by environment. *Molecular Ecology*,
757 23(23):5649–5662, 2014. ISSN 1365-294X. doi: 10.1111/mec.12938. URL <http://dx.doi.org/10.1111/mec.12938>.

- 759 [65] Andreas Weingessel and Kurt Hornik. Local PCA algorithms. *Neural Networks, IEEE*
 760 *Transactions on*, 11(6):1242–1250, 2000.
- 761 [66] Sewall Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):
 762 323–354, 1949. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1949.tb02451.x. URL <http://dx.doi.org/10.1111/j.1469-1809.1949.tb02451.x>.
- 764 [67] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price.
 765 Advantages and pitfalls in the application of mixed-model association methods. *Nat*
 766 *Genet*, 46(2):100–106, February 2014. ISSN 10614036. URL <http://dx.doi.org/10.1038/ng.2876>.

768 **A Choosing window length**

769 The choice of window length entails a balance between signal and noise. In very short
 770 windows, genealogies of the samples will only be represented by a few trees, so varia-
 771 tion between windows represents demographic noise rather than meaningful variation in
 772 patterns of relatedness. Longer windows generally have more distinct trees (and SNPs), al-
 773 lowing for less noisy estimation of local patterns of relatedness. However, to better resolve
 774 meaningful signal, i.e., differences in patterns of relatedness along the genome, we would
 775 like reasonably short windows.

776 Since we summarize patterns of relatedness using relative positions in the principal
 777 component maps, we quantify “noise” as the standard error of a sample’s position on PC1
 778 in a particular window, averaged across windows and samples, and “signal” as the standard
 779 deviation of the sample’s position on PC1 over all windows, averaged over samples. The
 780 definition of eigenvectors does not specify their sign, and so when comparing between
 781 windows we choose signs to best match each other: after choosing $PC1_1$, for instance,
 782 if u is the first eigenvector obtained from the covariance matrix for window j , then we
 783 next choose $PC1_j = \pm u$, where the sign is chosen according to which of $\|PC1_1 - u\|$ or
 784 $\|PC1_1 + u\|$ is smaller.

785 After doing this, the mean variance across windows is

$$\sigma_{\text{signal}}^2 = \frac{1}{N} \sum_{j=1}^N \frac{1}{L} \sum_{i=1}^L (PC1_{ij} - \overline{PC1}_j)^2,$$

786 where $PC1_{ij}$ is the position of the i^{th} individual on $PC1$ in window j , and $\overline{PC1}_j =$
 787 $(1/N) \sum_{j=1}^N PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jack-
 788 knife [9, 20]: we divide the j^{th} window into 10 equal-sized pieces, and let $PC1_{ijk}$ denote
 789 the first principal component of this region found after removing the k^{th} piece; then the

790 estimate of the squared standard error is $\sigma_{ij}^2 = \frac{9}{10} \sum_{k=1}^{10} (PC1_{ij,k} - \frac{1}{10} \sum_{\ell=1}^{10} PC1_{ij,\ell})^2$. Av-
791 eraging over samples and windows,

$$\sigma_{\text{noise}}^2 = \frac{1}{N} \sum_{j=1}^N \frac{1}{L} \sum_{i=1}^L \sigma_{ij}^2.$$

792 For the main analysis, we defined windows to each consist of the same number of neig-
793 boring SNPs, and calculated σ_{signal}^2 and σ_{noise}^2 for a range of window sizes (i.e., numbers
794 of SNPs). For our main results we chose the smallest window for which σ_{signal}^2 was con-
795 sistent larger than σ_{noise}^2 (but checked other sizes); the values for various window sizes
796 across *Drosophila* chromosomes are shown in Table S1. In the cases we examined, we found
797 nearly identical results after varying window size, and choosing windows to be of the same
798 physical length (in bp) rather than in numbers of SNPs.

chrom. arm		window length (SNPs)				
		100	500	1,000	10,000	100,000
2L	σ_{noise}^2	2.05	1.64	1.18	0.17	0.04
	σ_{signal}^2	2.76	2.69	2.23	0.68	0.31
2R	σ_{noise}^2	2.18	1.92	1.63	0.58	0.13
	σ_{signal}^2	2.78	2.70	2.65	2.31	1.82
3L	σ_{noise}^2	2.08	2.00	1.64	0.73	0.25
	σ_{signal}^2	2.60	2.52	2.40	1.68	1.89
3R	σ_{noise}^2	1.95	1.76	1.44	0.59	0.20
	σ_{signal}^2	2.58	2.51	2.44	1.96	1.40
X	σ_{noise}^2	2.48	2.04	1.54	1.62	0.17
	σ_{signal}^2	2.61	2.43	2.30	0.32	1.14

Table S1: Measures of signal and noise, computed separately for each chromosome arm in the *Drosophila* dataset, at different window sizes. All values are multiplied by 1,000 (so typical variation is of order of 50% of the actual values). Starting at windows of 1,000 SNPs, the signal (variation of PC1 between windows) starts to be substantially larger than the noise (standard error of PC1 for each window).

B Simulations

800 We implemented two types of simulation: first, simple simulations of Gaussian “geno-
801 types” where the expectation of variation in “population structure” was clear; and next,
802 individual-based simulations with explicit genomes, using SLiM.

803 **B.1 Gaussian simulations**

804 We simulated genotypes at each locus independently, drawing each vector of genotypes from
805 a multivariate Gaussian distribution with zero mean and covariance matrix Σ . Sampled
806 individuals came from three populations, and each Σ_{ij} depends on which populations the
807 individuals i and j are in, as well as the location along the chromosome. There are three
808 population-level mean relatedness matrices along the genome, which apply to the first
809 quarter ($S^{(1)}$), the middle half ($S^{(2)}$), and the last quarter ($S^{(3)}$), respectively:

$$S^{(1)} = \begin{bmatrix} 0.75 & 0.25 & 0.0 \\ 0.25 & 0.75 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$
$$S^{(2)} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.75 & 0.25 \\ 0.0 & 0.25 & 0.75 \end{bmatrix}$$
$$S^{(3)} = \begin{bmatrix} 0.75 & 0.0 & 0.25 \\ 0.0 & 1.0 & 0.0 \\ 0.25 & 0.0 & 0.75 \end{bmatrix}$$

810 If individuals i and j are in populations $p(i)$ and $p(j)$ respectively, then the covariance
811 between their genotypes is $\Sigma_{ij} = S_{p(i),p(j)}$, using the appropriate S for that segment of the
812 genome. The variance of individual i 's genotype is $\Sigma_{ii} = S_{p(i),p(i)} + 0.1$.

813 We first created “genotypes” in this way with fifty individuals from each of the three
814 populations; running our method on a genome with 99 windows of 400 loci each produced
815 the first plot in Figure S19. These matrices are chosen so that the top two eigenvalues
816 Σ are the same (both 50.1), and so the ordering of the top two PCs is arbitrary. If our
817 method was sensitive to PC ordering, then half the windows in each region that have one
818 ordering would cluster with each other, separate from the other half.

819 We then marked each genotype in the first half of the chromosome as missing, inde-
820 pendently, with probability 1/2 and ran our method again, producing the second plot of
821 Figure S19. If our method was influenced by missing data, we would expect the first half
822 of the chromosome to separate from the second in the MDS plot.

823 **B.2 SLiM simulations**

824 Our SLiM simulations were constructed as follows. Individuals are diploid, and genomes
825 have a length of 153,520,244 bp. Recombination was either (a) flat, with a constant rate of
826 10^{-9} ; (b) according to the human female HapMap map for chromosome 7; or (c) constant
827 in each of seven equal-sized regions, beginning at 2.04×10^{-8} , descending by a factor of four
828 for three steps, and then ascending by a factor of four for three steps, so that the middle

829 seventh has the lowest recombination rate, and the outer two sevenths has a rage 64 times
830 higher. Selected mutations are introduced at a rate of 10^{-10} per bp per individual per
831 generation, and have selection coefficients drawn from a Gamma distribution with mean
832 0.005 and shape 2; each coefficient are either positive or negative with probabilities 1/30
833 and 29/30 respectively. Each simulation was run for 50,000 generations.

834 Each individual has a spatial position in the two-dimensional square of width $W =$
835 8. Each time step, each individual chooses the nearest other to mate with, producing a
836 random, Poisson distributed number of offspring with mean 1/3. Offspring are assigned
837 random spatial locations displaced from their parent's by a bivariate Gaussian with mean
838 zero and standard deviation $\sigma = 0.2$, reflected to stay within the habitat range.

839 Each individual survives to the next time step with probability equal to their fitness.
840 Fitness values are determined multiplicatively by the effects of each mutation, but are
841 multiplied by an additional factor determined by the local density of individuals. This
842 factor is equal to $\rho/(1 + C)$, where $\rho = 2\pi K\sigma^2$ is the carrying capacity per circle of
843 radius σ ; $K = 100$ is the mean equilibrium population density; and C is the sum of a
844 Gaussian kernel with standard deviation $\sigma = 0.1$ between the focal individual and all other
845 individuals within distance 3σ . To avoid edge effects, fitnesses are further multiplied by
846 $\min(1, z)$, where z is the distance to the nearest boundary. This produces populations that
847 fluctuate at equilibrium around 6,000 individuals in total, fairly evenly spread across the
848 square.

849 In one additional simulation, we modified fitnesses by multiplying the selective effect of
850 each allele in each individual by multiplying it by $2x/W - 1$, where x is the x coordinate
851 of the individual. This makes the effect of each allele opposite on the left and on the right,
852 and neutral in the middle, and leads to a moderate number of balanced polymorphisms.

853 **C Supplementary Tables**

854 **D Supplementary Figures**

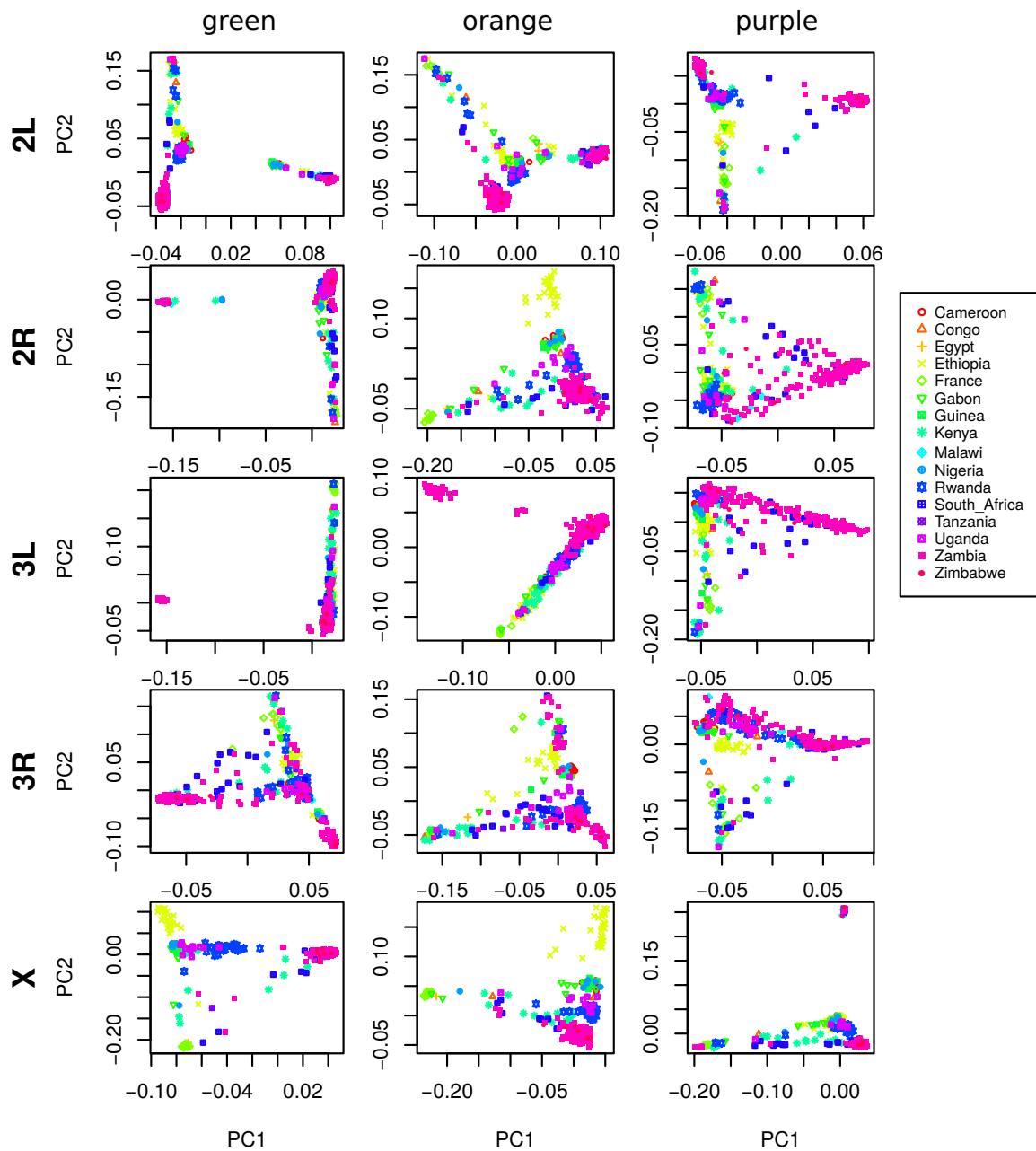


Figure S1: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange windows; and third is for purple windows.

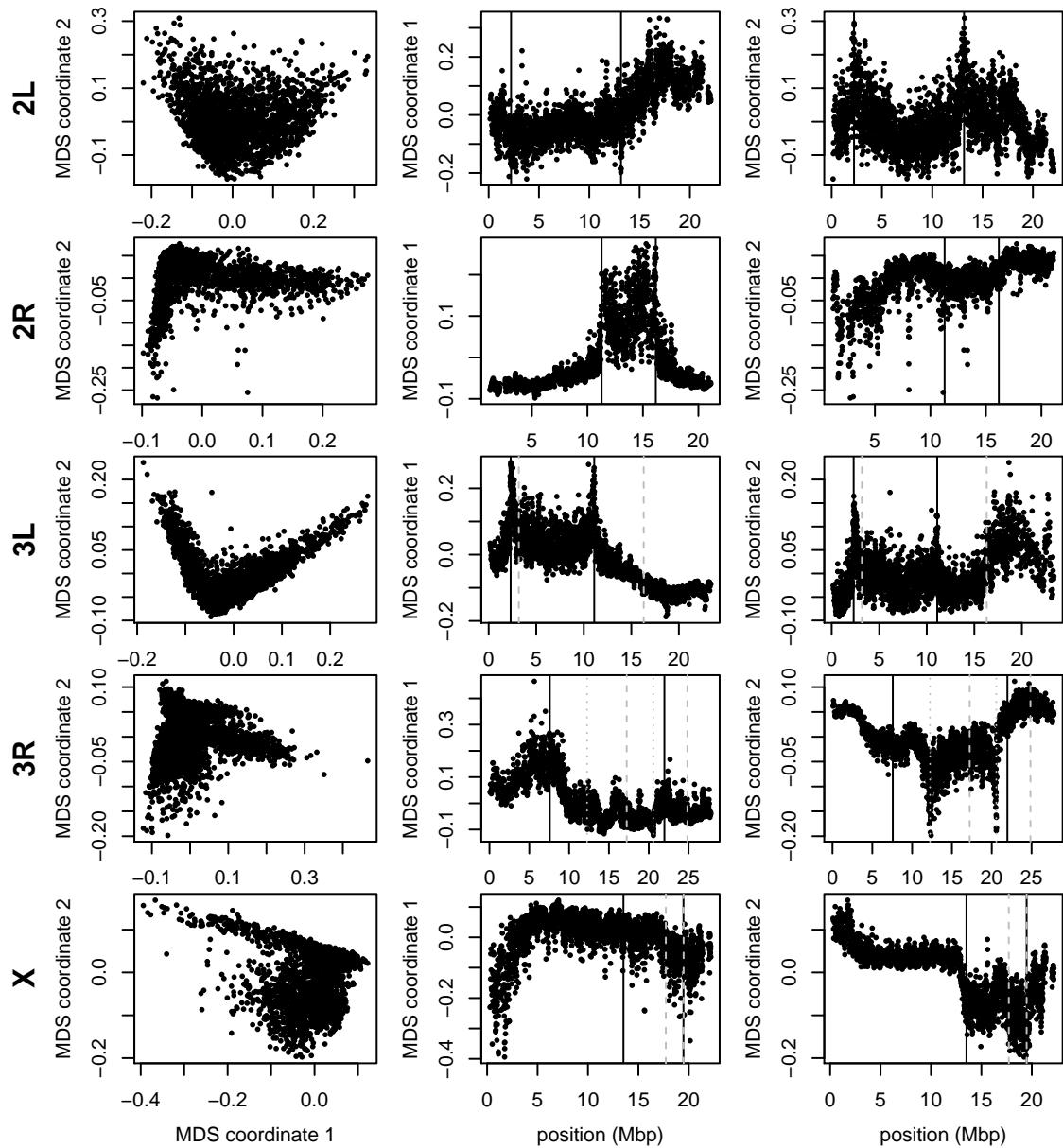


Figure S2: MDS visualizations for each chromosome arm of *Drosophila melanogaster*, as in Figure 2, except that the method was run using five PCs ($k = 5$) instead of two.

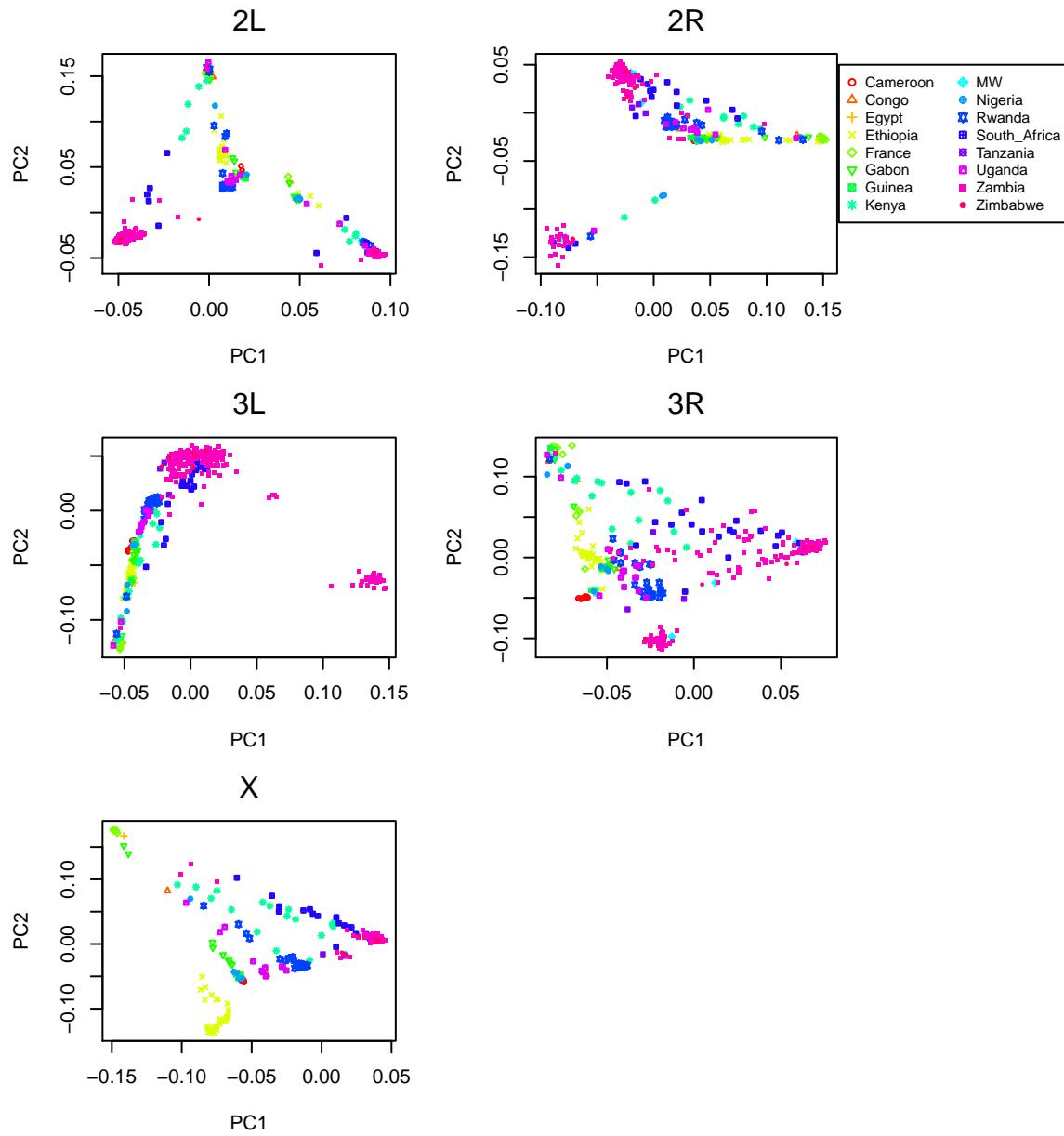


Figure S3: PCA plots for chromosome arms 2L, 2R, 3L, 3R and X of the *Drosophila melanogaster* dataset.

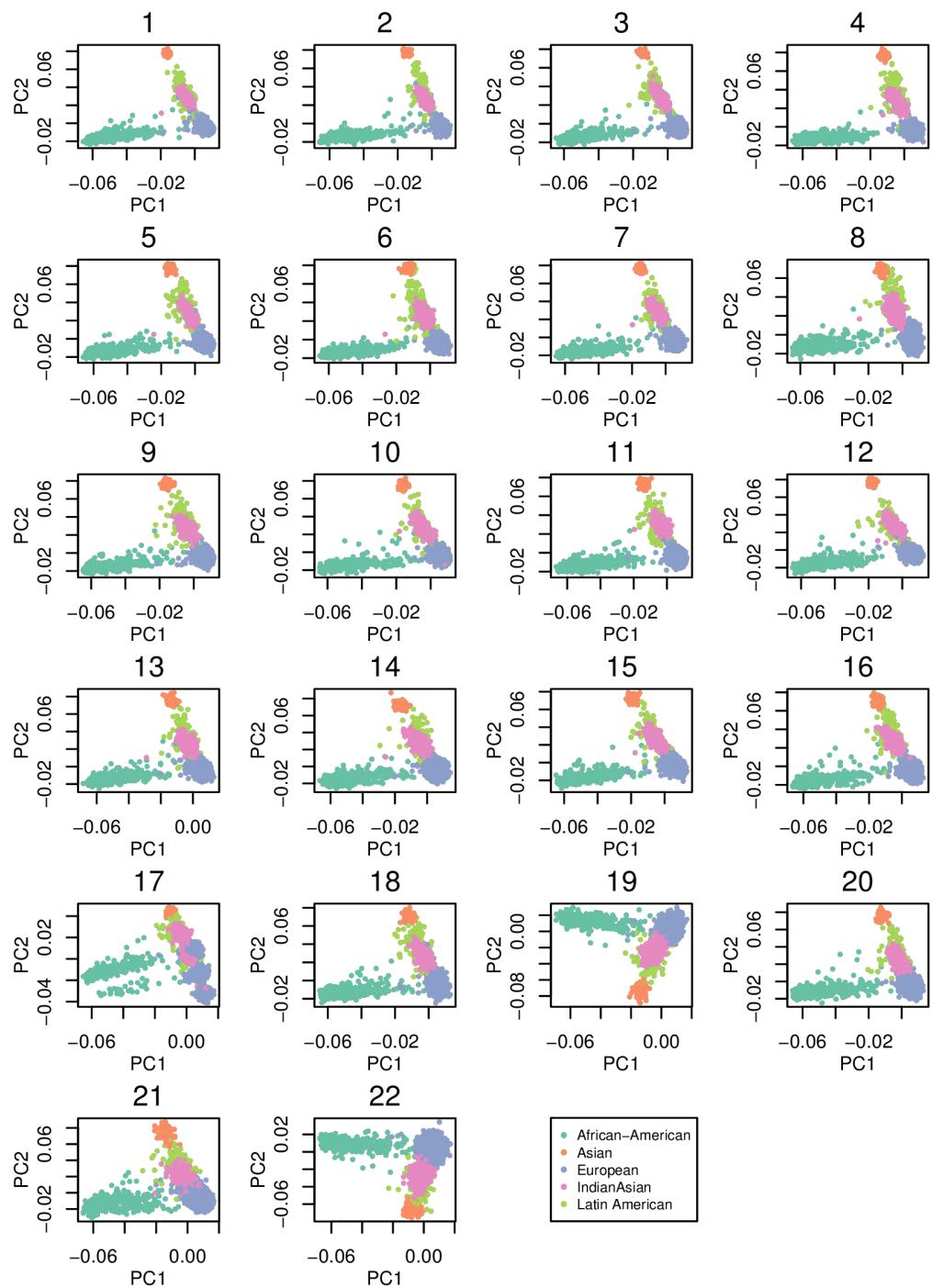


Figure S4: PCA plots for all 22 human autosomes from the POPRES data.

	10000 SNPs MDS1	10000 SNPs 2 PCs	1000 SNPs 2 PCs	10000 SNPs 5 PCs	100000bp 2 PCs	10000bp 2 PCs
10000 SNPs, 2 PCs		1.00	0.87	0.96	0.90	0.88
1000 SNPs, 2 PCs		0.68	1.00	0.73	0.68	0.94
10000 SNPs, 5 PCs		0.96	0.92	1.00	0.88	0.93
100000bp, 2 PCs		0.90	0.87	0.88	1.00	0.87
10000bp, 2 PCs		0.68	0.93	0.72	0.67	1.00
MDS2						
10000 SNPs, 2 PCs		1.00	0.54	0.93	0.87	0.56
1000 SNPs, 2 PCs		0.82	1.00	0.76	0.83	0.92
10000 SNPs, 5 PCs		0.93	0.50	1.00	0.83	0.52
100000bp, 2 PCs		0.87	0.59	0.84	1.00	0.58
10000bp, 2 PCs		0.83	0.92	0.77	0.84	1.00

Table S2: Correlations between MDS coordinates of genomic regions between runs with different parameter values. To produce these, we first ran the algorithm with the specified window size and number of PCs (k in equation (1)) on the full *Medicago truncatula* dataset. Then to obtain the correlation between results obtained from parameters A in the row of the matrix above and parameters B in the column of the matrix above, we mapped the windows of B to those of A by averaging MDS coordinates of any windows of B whose midpoints lay in the corresponding window of A; we then computed the correlation between the MDS coordinates of A and the averaged MDS coordinates of B. This is not a symmetric operation, so these matrices are not symmetric. As expected, parameter values with smaller windows produce noisier estimates, but plots of MDS values along the genome are visually very similar.

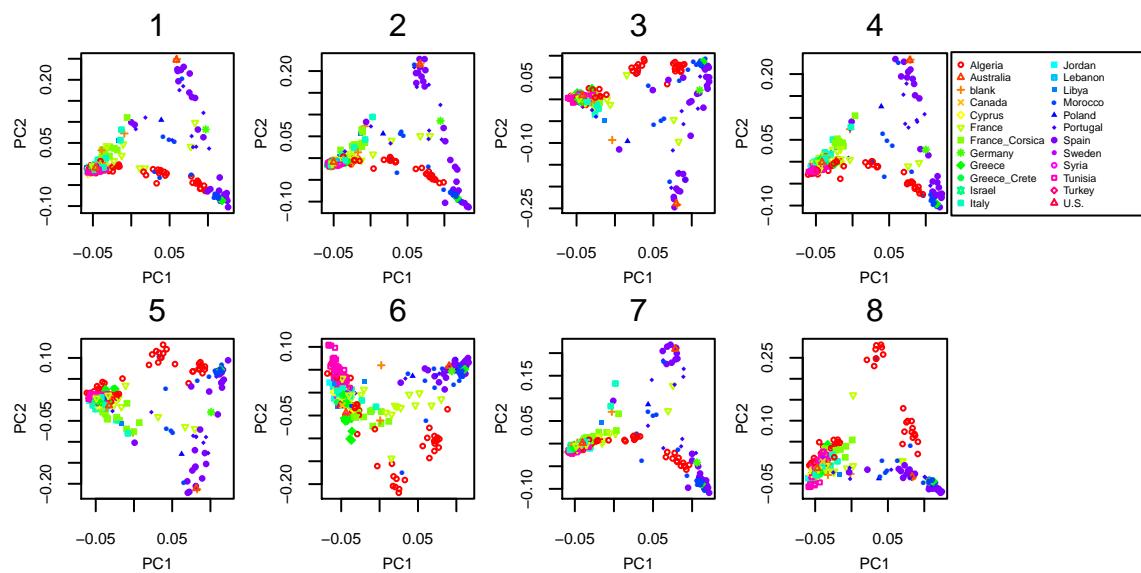


Figure S5: PCA plots for all 8 chromosomes in the *Medicago truncatula* dataset.

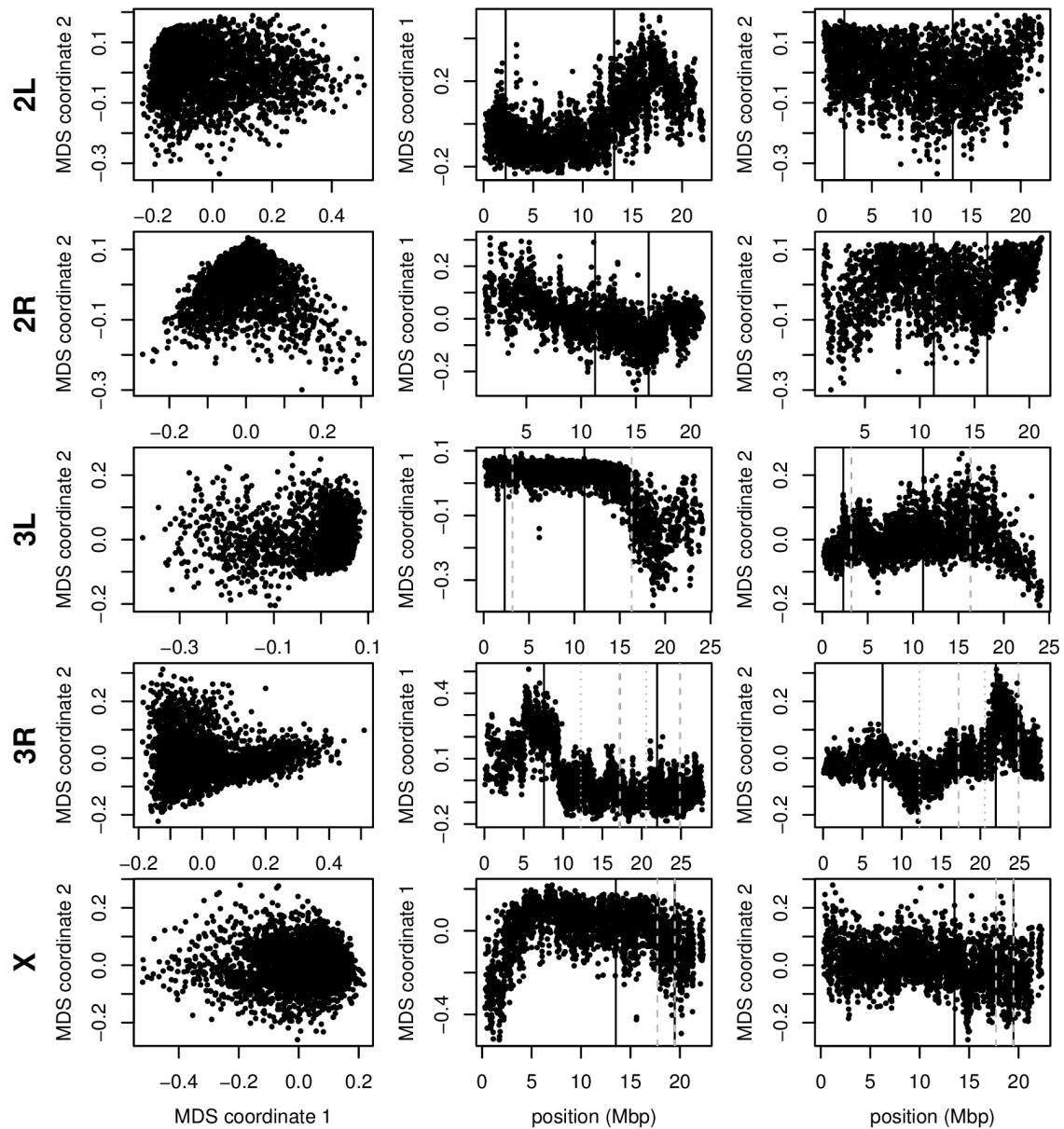


Figure S6: Variation in structure for windows of 1,000 SNPs across *Drosophila melanogaster* chromosome arms: without inversions. As in Figure 2, but after omitting for each chromosome arm individuals carrying the less frequent orientation of any inversions on that chromosome arm. The values differ from those in 4 in the window size used and that some MDS values were inverted (but relative orientation is meaningless as chromosome arms were run separately, unlike for *Medicago*). In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosome arms. Colors are consistent for plots in each row. Vertical lines show the breakpoints of known polymorphic inversions.

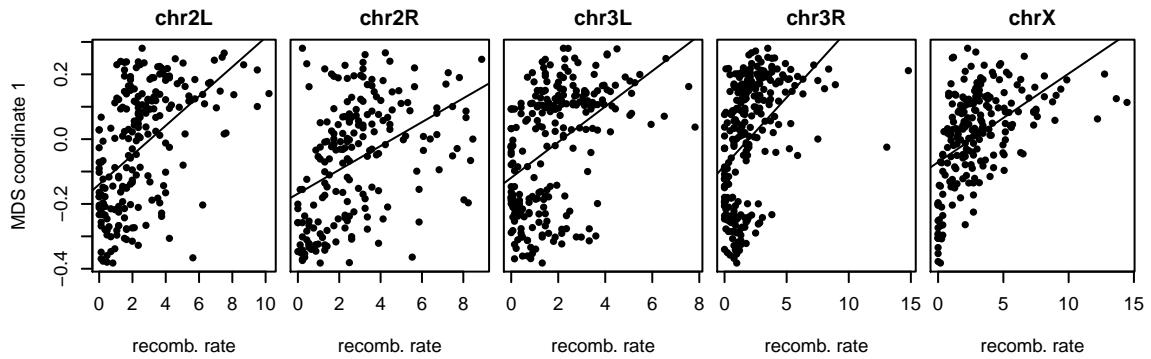


Figure S7: Recombination rate, and the effects of population structure for *Drosophila melanogaster*: this shows the first MDS coordinate and recombination rate (in cM/Mbp), as in Figure 4, against each other. Since the windows underlying estimates of Figure 4 do not coincide, to obtain correlations we divided the genome into 100Kbp bins, and for each variable (recombination rate and MDS coordinate 1) averaged the values of each overlapping bin with weight proportional to the proportion of overlap. The correlation coefficient and p -values for each linear regression are as follows: 2L: correlation = 0.52, r^2 = 0.27; 2R: correlation = 0.43, r^2 = 0.18; 3L: correlation = 0.47, r^2 = 0.21; 3R: correlation = 0.46, r^2 = 0.21; X: correlation = 0.50, r^2 = 0.24.

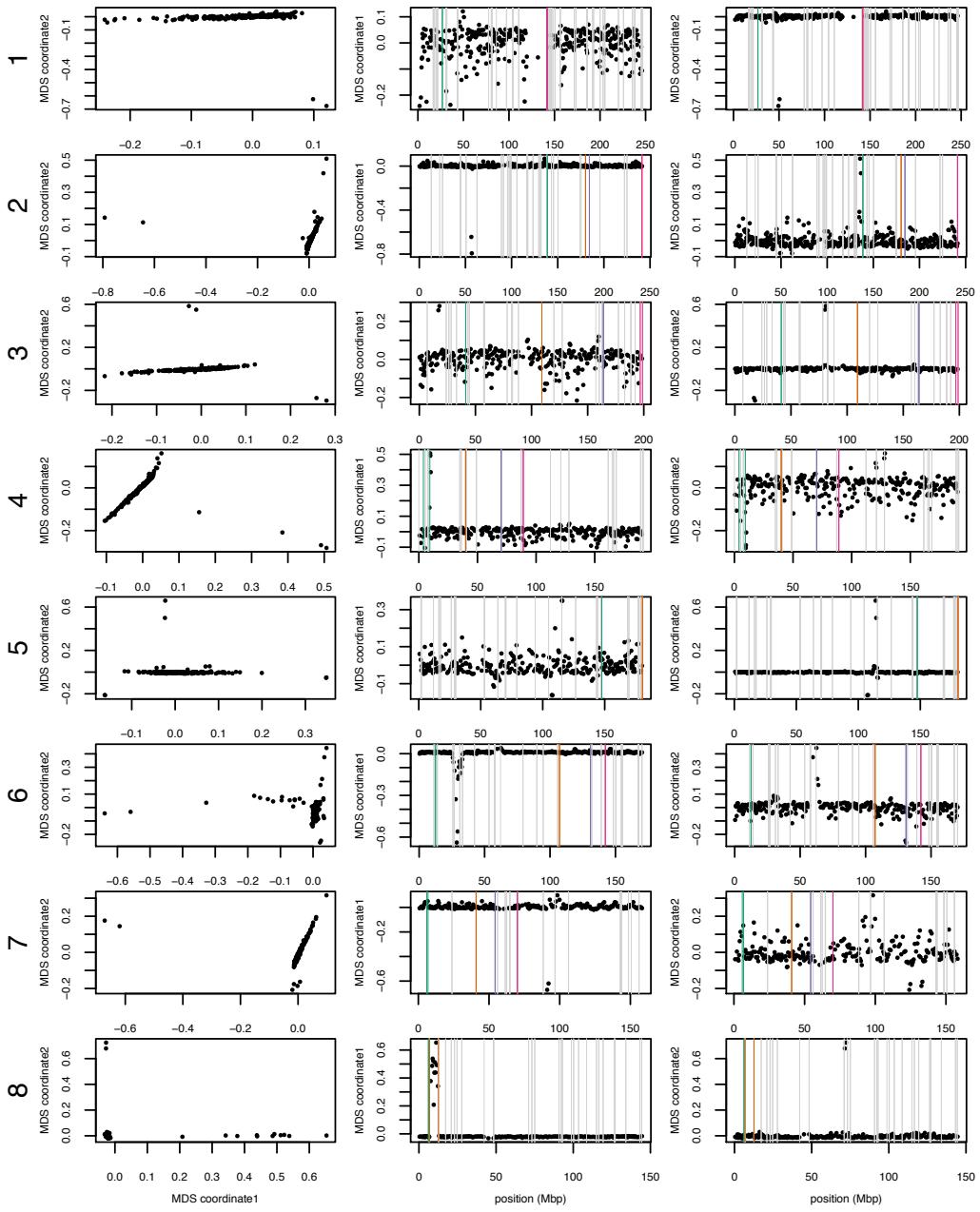


Figure S8: MDS plots for human chromosomes 1-8. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosomes. Colorful vertical lines show the breakpoints of known valid inversions, while grey vertical lines show the breakpoints of predicted inversions.

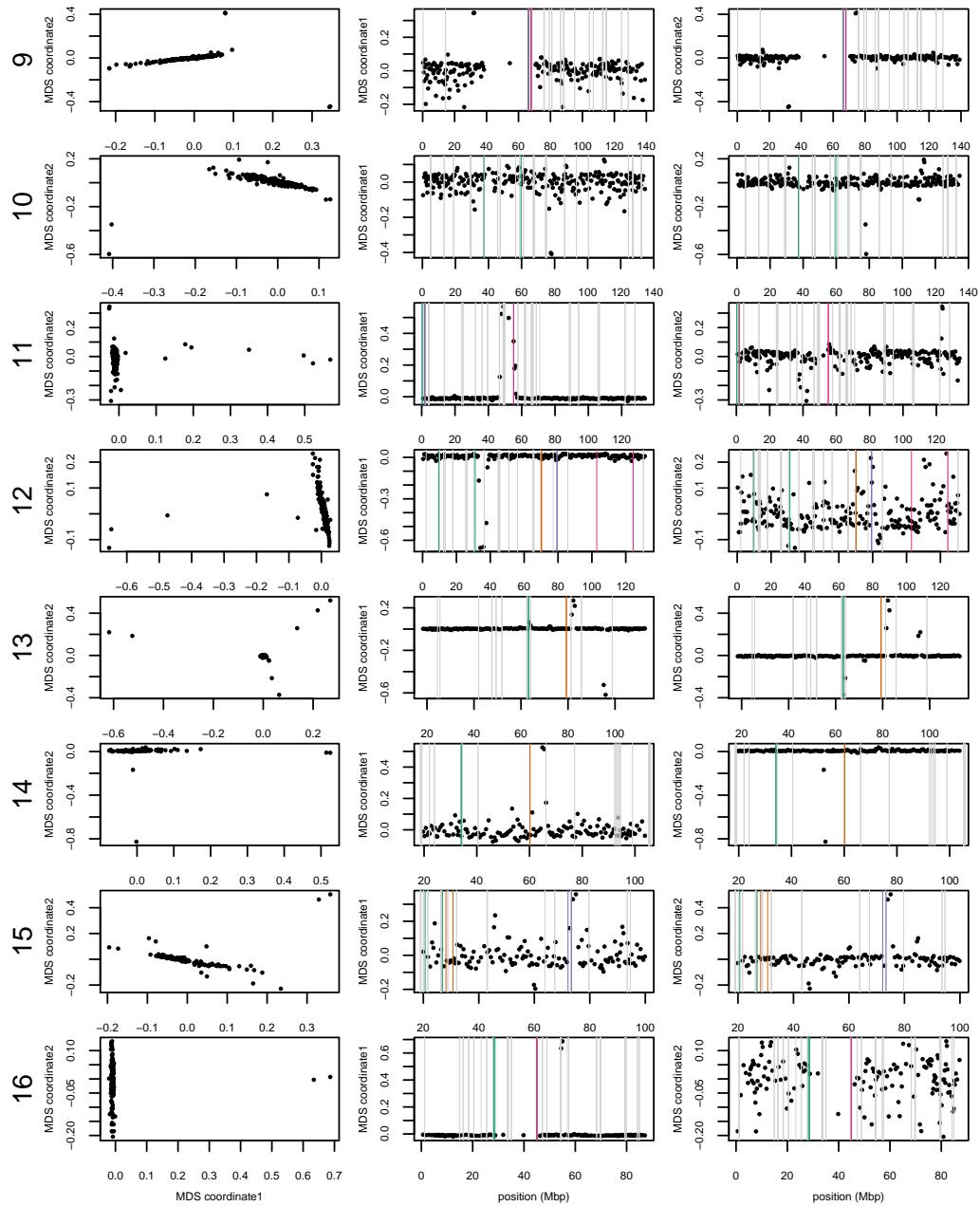


Figure S9: MDS plots for human chromosomes 9-16, as in Supplemental Figure S8.

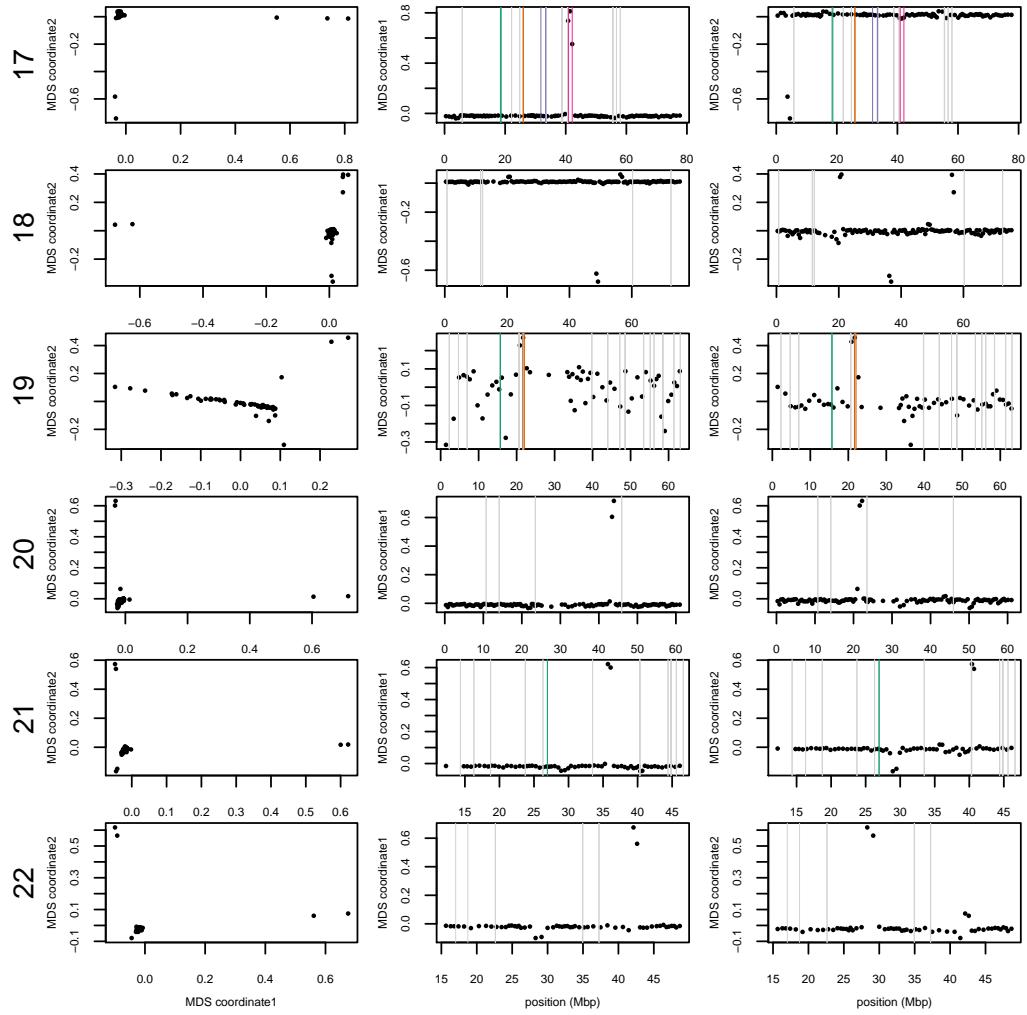


Figure S10: MDS plots for human chromosomes 17-22, as in Supplemental Figure S8.

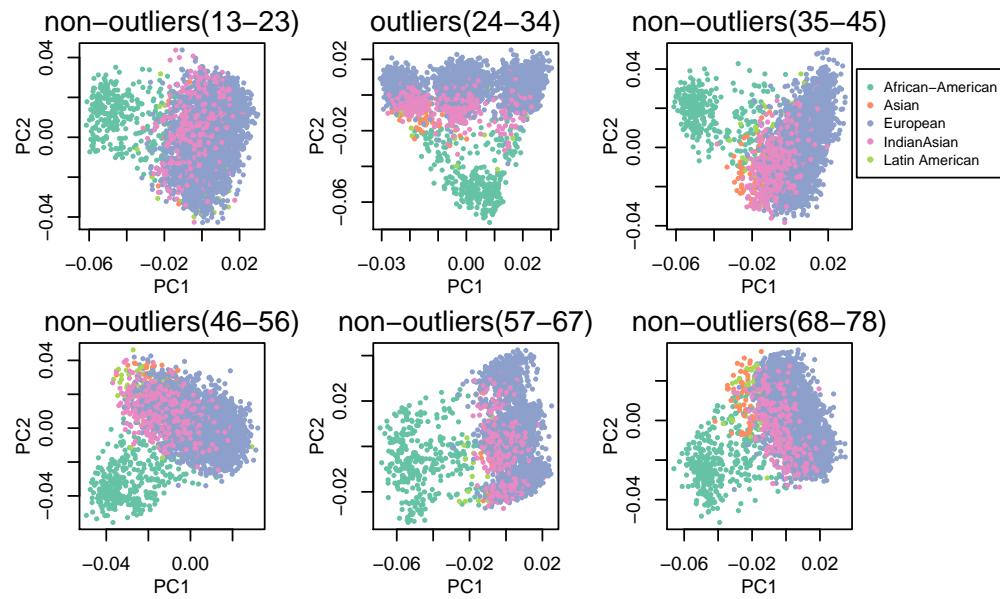


Figure S11: Comparison of PCA figures within outlying windows (center column) and flanking non-outlying windows (left and right columns) for the two windows having outlying MDS scores on chromosome 8.

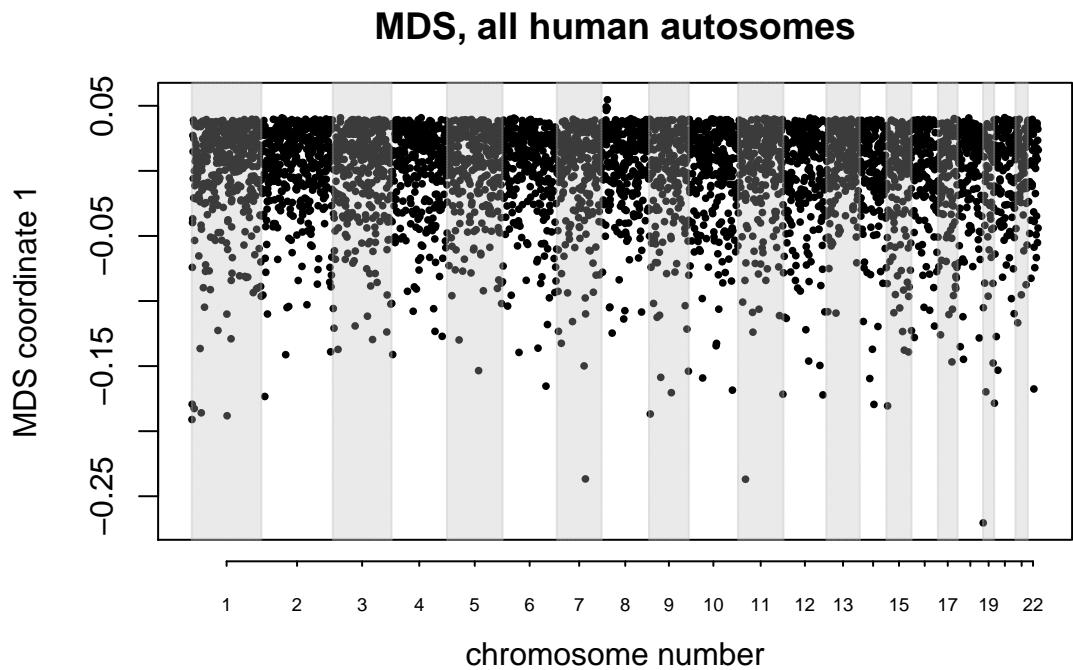


Figure S12: MDS visualization of variation in the effects of population structure amongst windows across *all* human autosomes simultaneously. The small group of windows with positive MDS values lie around the inversion at 8p23.

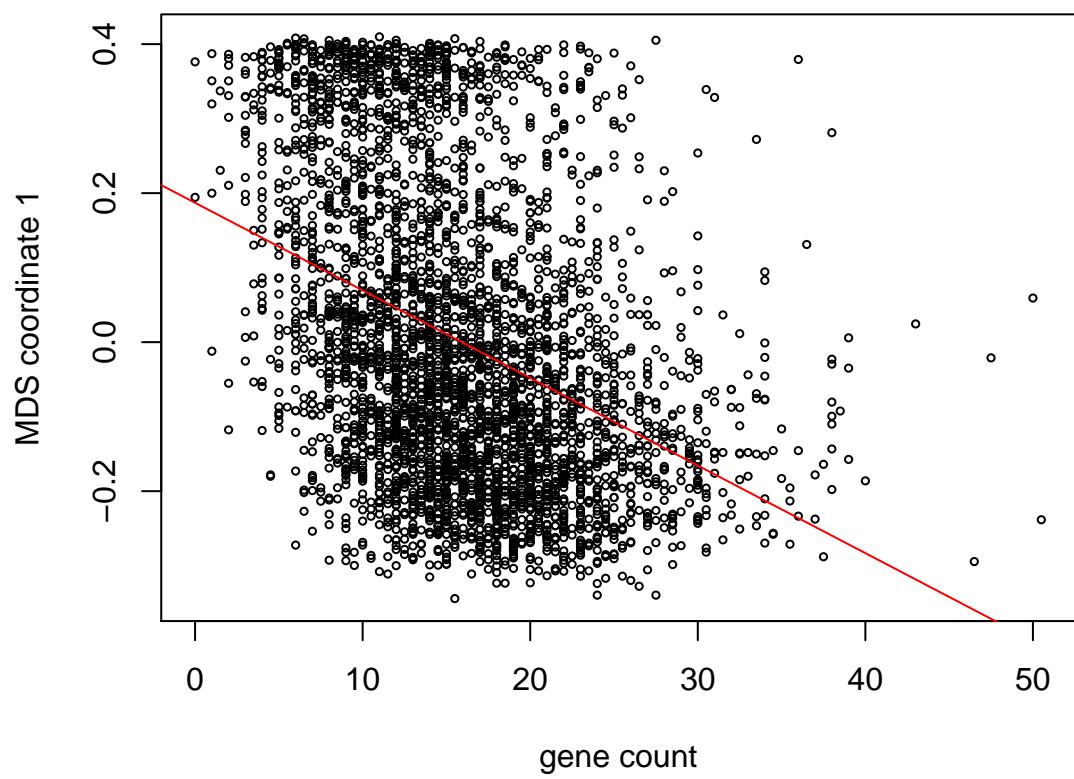


Figure S13: First MDS coordinate against gene density for all 8 chromosomes of *M. truncatula*. The first MDS coordinate is significantly correlated with gene count ($r = 0.149$, $p = 2.2 \times 10^{-16}$).

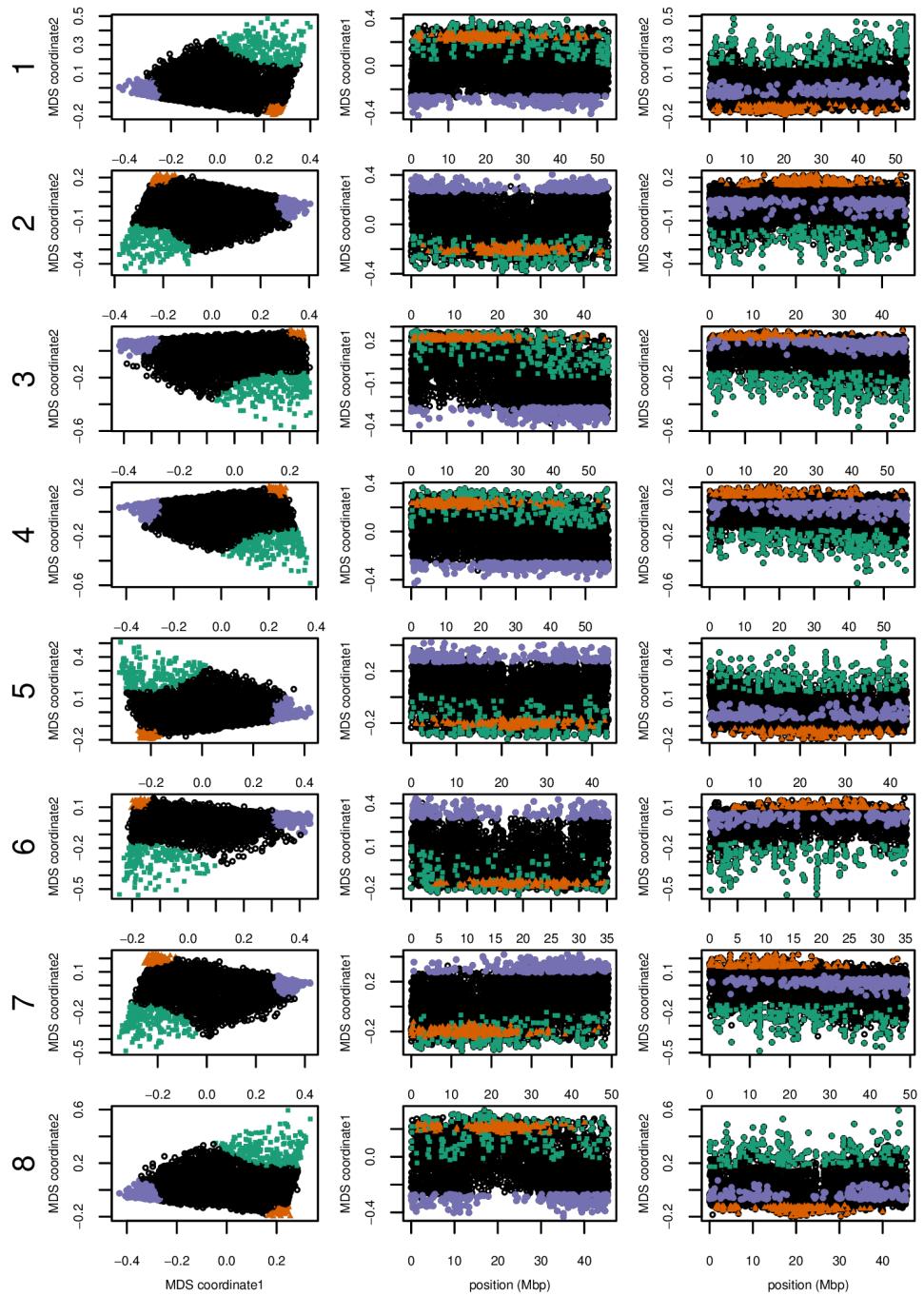


Figure S14: MDS visualizations of the effects of population structure for all 8 chromosomes of the *Medicago truncatula* data, using windows of 10^4 SNPs.

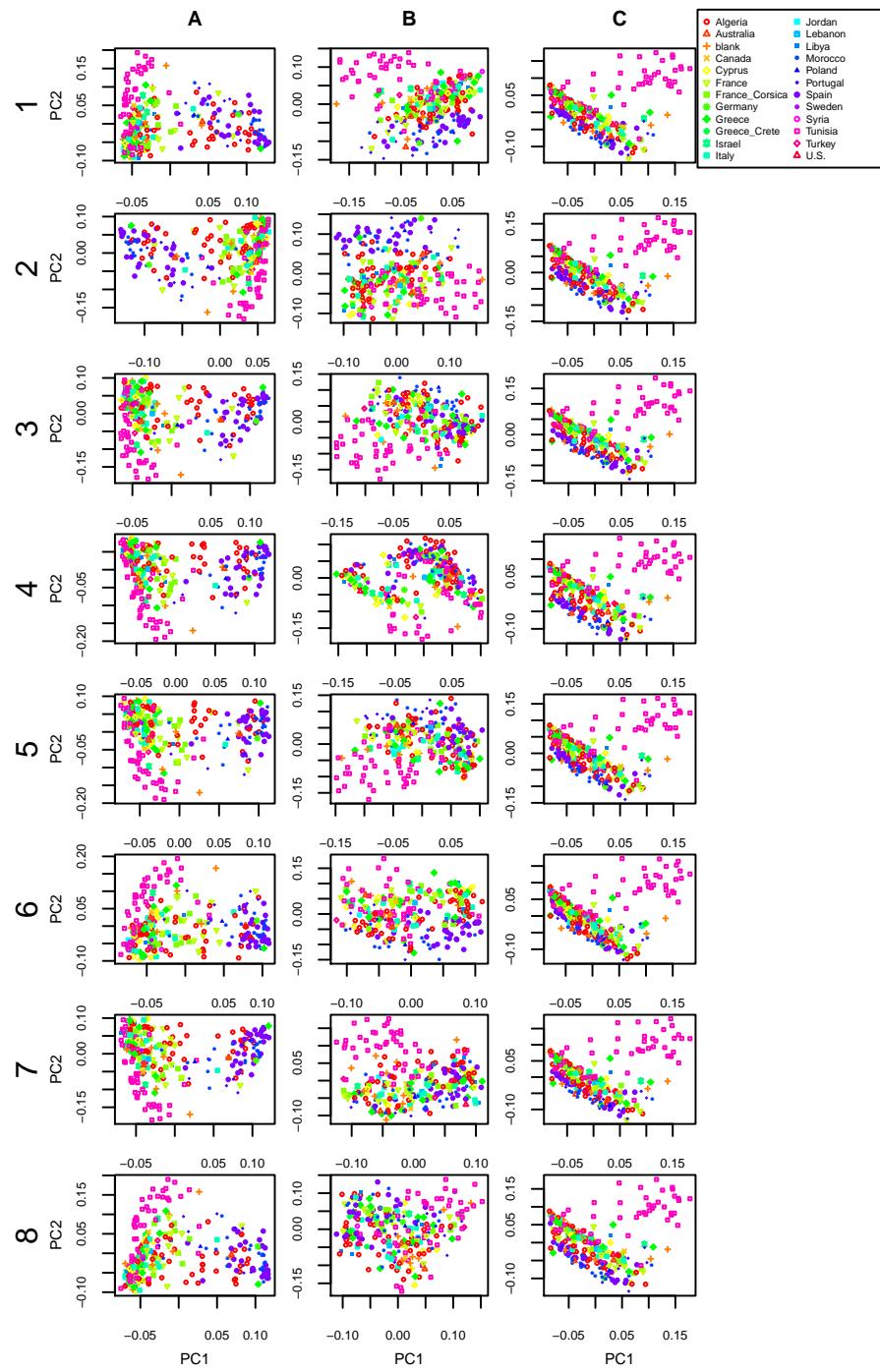


Figure S15: PCA plots for regions colored in Figure S14 on all 8 chromosomes of *Medicago truncatula*: (A) green, (B) orange, and (C) purple.

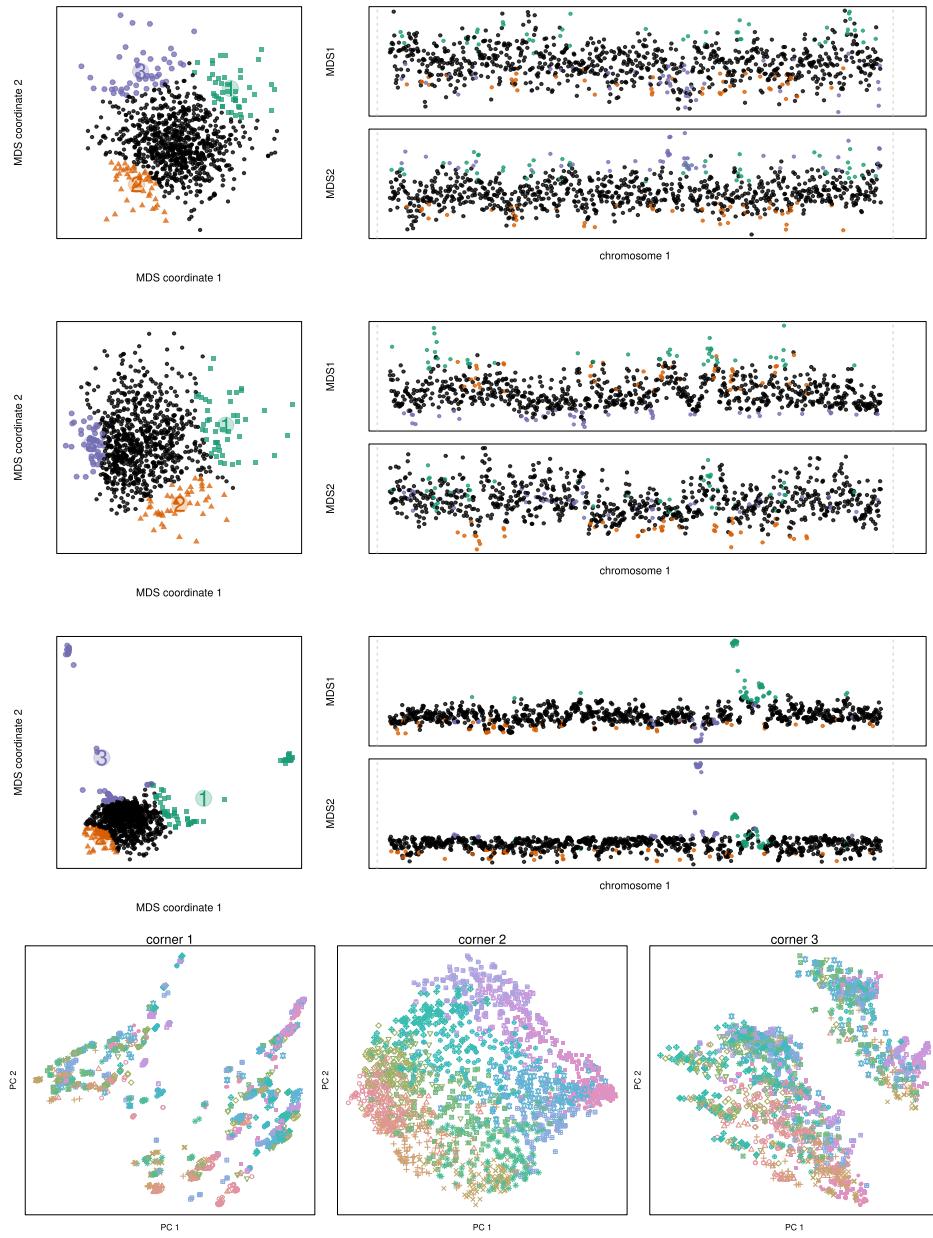


Figure S16: MDS visualizations of the results of individual-based simulations using SLiM (see Appendix B.2 for details). All simulations are neutral, and recombination is: **(top)** constant; **(top middle)** varies stepwise by factors of two in seven equal-length segments, with highest rates on the ends, so the middle segment has a recombination rate 64 times lower than the ends; **(bottom middle)** according to the HapMap human female chromosome 7 map. The **bottom** figure shows PCA maps corresponding to the three colored windows of the last (HapMap) situation; the ⁴⁹ outlying regions are long regions of low recombination rate, so that region can be dominated by a few correlated trees, similar to an inversion.

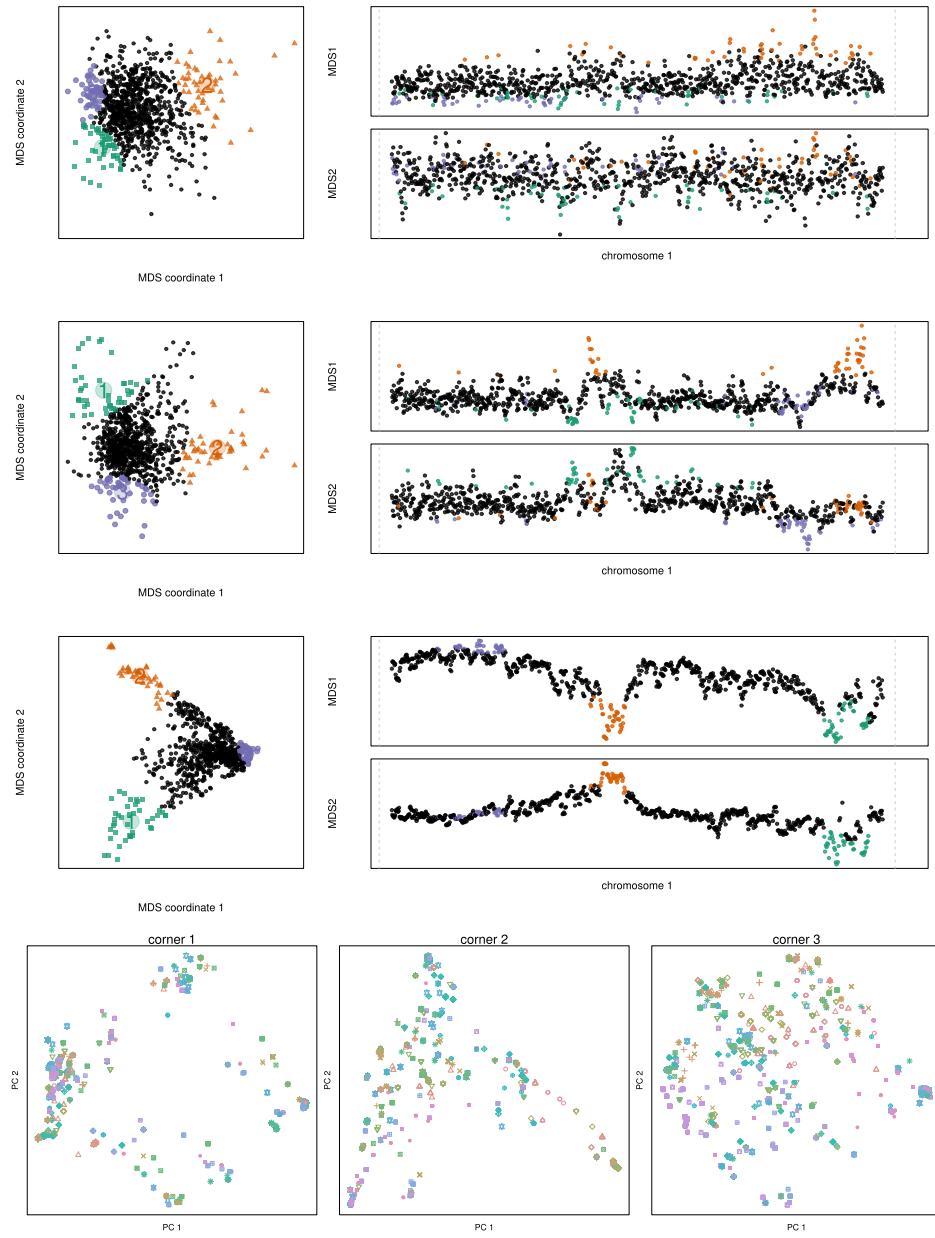


Figure S17: MDS visualizations of the results of individual-based simulations using SLiM (see Appendix B.2 for details). All simulations incorporate linked selection by allowing selected mutations to appear in the same two regions of the genome: the one-sixth of the genome immediately before the halfway point, and the last one-sixth of the genome. **(top)** Constant recombination rate. **(top middle)** Stepwise varying recombination rate (as described in Figure S16). **(bottom middle)** Constant recombination rate with spatially varying effects of selection. **(bottom)** PCA⁵⁰ plots corresponding to the highlighted corners of the last MDS visualization, showing how spatially varying linked selection has affected patterns of relatedness.

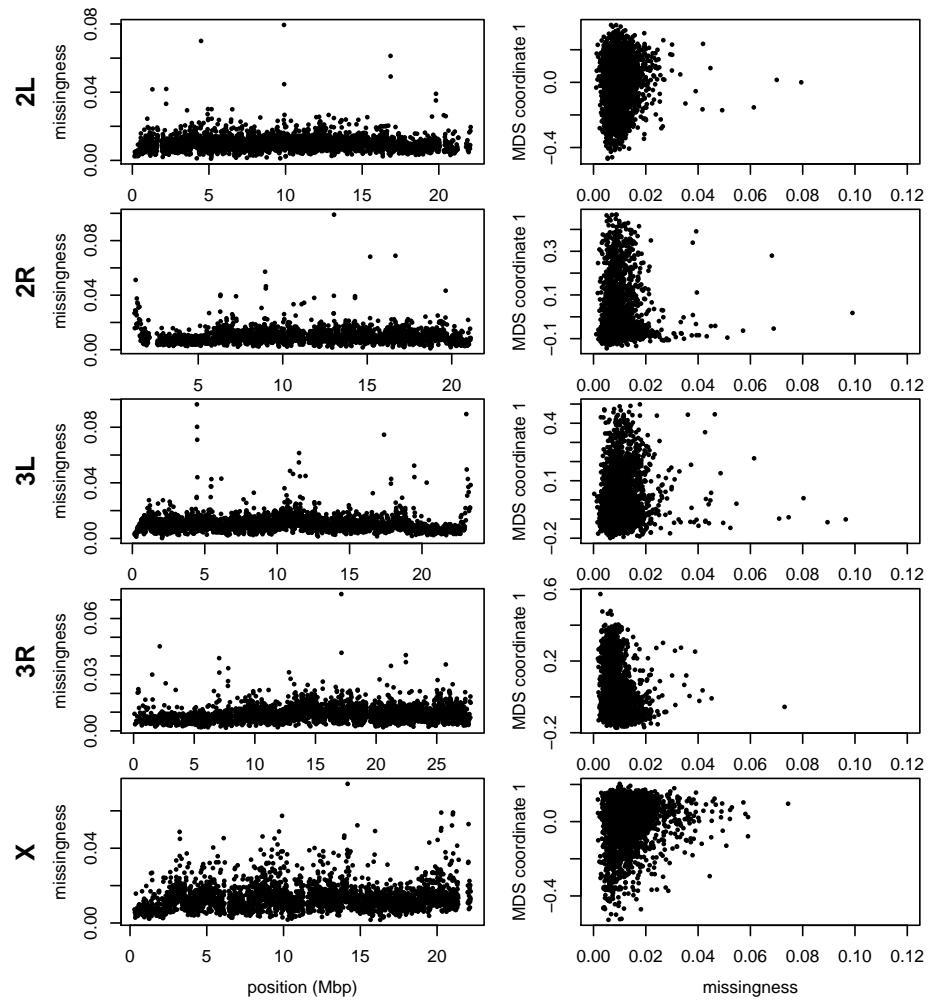


Figure S18: The proportion of data in each window that are missing, compared to the value of the first MDS coordinate for the *Drosophila melanogaster* data from Figure 2.

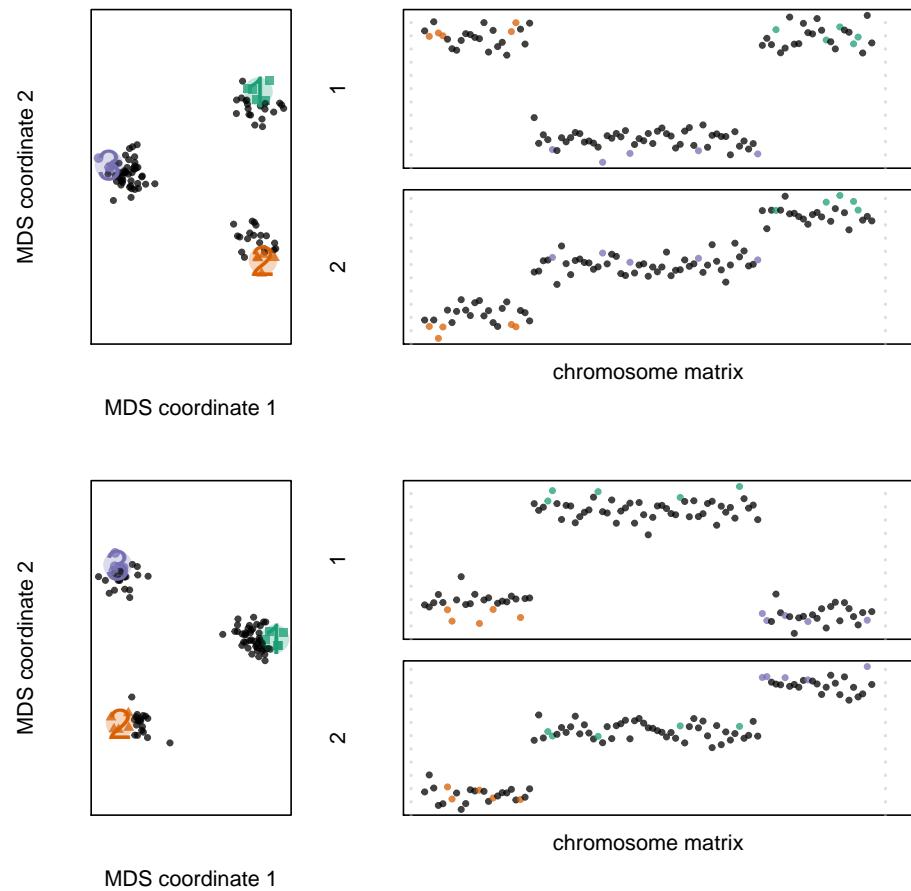


Figure S19: MDS visualizations of the Gaussian genotypes described in Appendix B.1, for 50 individuals from each of three populations. **(top)** The first quarter, middle half, and final quarter of the chromosome each have different population structure, as expected, despite the possibility for PC switching within each. **(bottom)** The same picture results even after marking a random 50% of the genotypes in the first half of the chromosome as missing.

Resubmission Cover Letter
Genetics

Han Li
and Peter Ralph
August 22, 2018

To the Editor(s) –

We are pleased to submit a revision of our manuscript,
Sincerely,

Han Li and Peter Ralph

Reviewer AE:

Please understand that incremental changes will not be sufficient. Adding simulations to strengthen key claims will be necessary, particularly addressing the impacts of mutation rate and recombination rate variation with more depth, the concern regarding PC switching (Reviewer 1), and the concern regarding the impacts of variation in missingness by sub-population (Reviewer 2).

Thanks for the positive feedback and the useful suggestions. We agree that more extensive exploration using simulations would help bolster understanding of the method, and have now done so. This took a substantial amount of work, because genome-scale forwards-time simulations with a large number of loci under selection is at or beyond the current limits of computation, depending on the number of individuals simulated.

Reviewer 1:

The paper is generally well written and clear; it addresses an important problem, and clearly makes some progress on it. However, it suffers from having no grounding in either theory or empirical demonstration that it really can find the structures that are claimed. I find the arguments that it finds inversions compelling, though not watertight, and I am not yet convinced that it is finding ubiquitous background selection. To make this claim, significant extra work is required.

In short, the approach is interesting but not sufficiently explored to produce compelling evidence for the implications that are claimed. Putting a large amount of effort into simulations may alleviate these concerns somewhat.

Specific points: What does this method find? I'm concerned about: (a) variation in the recombination rate and (b) variation in the mutation rate, creating spurious structure.

The first possibility is that massively varying information quantity within windows could lead to a small number of such windows having their orientation reversed: that is, PC1 becomes PC2 and vice versa. (Or PC2 and PC3 could switch). This would lead to such windows having unusual properties and hence appearing as evidence of an inversion.

I do agree with the authors that significant outliers would be found at inversions. However, even if the PC switching does not occur, or the model could handle it, the evidence for selection is weaker. If the two types of variation described above exist, with no selection, I would still expect a “continuous triangle” of results (as seen left of Fig 2, top left of Fig 6) with extrema described by windows

with the most information, and points placed at different extremum having low recombination rate (because by chance, these will get an approximately fixed local tree, corresponding on average to the genome-wide population structure).

Addressing this is likely quite hard, though the authors may be able to think of something that separates these effects from selection.

(1.1) ... (a) variation in the recombination rate and (b) variation in the mutation rate ... creating spurious structure.

Reply: Since we only look at SNPs, a 1Kb window with mutation rate μ and recombination rate ρ is equivalent to a 10Kb window with mutation rate $\mu/10$ and recombination rate $\rho/10$, and the ratio μ/ρ determines the “signal-to-noise” ratio. For this reason, varying recombination rate in simulations, and varying window size and type in analysis suffices to check for the effects of both recombination and mutation rate. We have now added simulation tests showing that large-scale variation in recombination rate does not create spurious structure (p. 9, l. 255); except in the presence of long regions of low recombination, as expected (p. 9, l. 259). The point is addressed by comparing results with windows of different types – windows of equal length in bp (or in SNPs) have different lengths in cm; since these different choices show nearly identical patterns in all cases we have examined, recombination rate variation cannot be driving the results (e.g., compare Figures 2 and Supplemental Figure [XXX](#)).

(1.2) PC switching ... could lead to a small number of such windows having their orientation reversed: that is, PC1 becomes PC2 and vice versa. (Or PC2 and PC3 could switch).

Reply: This is a natural concern. However, the only point at which we compare PCs in a way that could be sensitive to ordering is in determining the window size – in computing the distance between windows we use a measure which is invariant under ordering. We have made this more clear by moving the note about flipping signs of PCs to the appendix on window choice (p. 29, l. 784) and added more explicit notes about this to (p. 4, l. 134) and (p. 4, l. 141). Furthermore, one of our simulation tests is explicitly designed to test for this effect. (p. 31, l. 818)

(1.3) p6 “here, we use $k=2\dots$ ” - you have to show that $k > 2$ is the same.

Reply: We have added a table including correlations between MDS values between $k = 2$ and $k = 5$ for the Medicago data (Table S2), and a representative figure (for the Drosophila data, $k = 5$), see Figure S2. (p. 10, l. 282)

(1.4) p15 “We also found nearly identical results when choosing shorter windows of 1,000 SNPs” - again, show this.

Reply: The same table includes correlations between MDS values for different window lengths (Table S2). (p. 16, l. 365)

(1.5) p15 “or choosing windows of equal length in base pairs rather than SNPs” - once again.

Reply: Again, we address this in Table S2. (p. 16, l. 365)

(1.6) Using 2 PCs is common practice: only if this is the end of an analysis and the PCA was done for visualisation. Here you are using it for something so should keep all the relevant PCs.

Reply: This is a good point; the question is which the “relevant” PCs are. Novembre and Stephens [50] showed that under isolation by distance, the top two PCs should reflect the two-dimensional nature of the range, and higher PCs are generally much less interpretable; we used $k = 2$ with this in mind. We have changed this sentence (p. 16, l. 365).

(1.7) I’m surprised that PCAdmix isn’t referenced. It is using a very similar method, albeit with different goals. In particular, the approach of placing all points into a single, genome-wide PC space solves many of the problems that this approach has (though I agree there may be benefits to the approach described here)

Reply: Good point: we now reference this work and discuss its similarity. (p. 3, l. 79)

Reviewer 2:

This is an interesting and well written paper. It was a pleasant read. I have three main general comments:

Thanks! We’re glad you enjoyed it.

(2.1) Related work: The authors provide an introduction of the main concepts, as well as some intuition of what the method is doing and how, but I found comparison to previous approaches to be somewhat missing. To some extent, this is due to the fact that the main goal of their analysis is somewhat vaguely “finding heterogeneity”, which leads to the applications of detecting chromosomal inversions and evidence for background selection. It would help to have a well defined set of hypotheses, test the method’s accuracy using simulation (see next comment), and compare to previous efforts in similar domains.

Reply: First: we think that “finding heterogeneity” is in fact a well-defined goal, although it was not that well-defined in the paper; we have hopefully improved on this in the

Introduction (p. 3, l. 103). Expanding a bit more: We strongly agree that methods that seek to test well-defined hypotheses are extremely useful and powerful. We also feel that methods for visualization and exploration are also useful – a primary example here being PCA. If PCA is useful – and we think that it is – then it should be important to also know how much the thing that PCA is summarizing varies along the genome, in the same way that knowing the mean of some quantity in a population is only of limited usefulness without also knowing the corresponding population variance.

(2.2) Validation: *In several occasions, the authors seem to introduce a potential problem in their approach, and provide a solution to it. This is generally rather intuitive, but it would really help to have simulations of some sort to show that the issue arises and leads to a problem, and that their approach does address the specific problem.*

Reply: We have added a number of simulations testing various aspects of the method, and have tried to incorporate the results in the paper without cluttering it up with the results of sanity checks. See sections 2.4 and 3.1; in particular the Gaussian simulations described in appendix B.1.

(2.3) *The use of weighted PCA to cope with unbalanced sample size could be better demonstrated. Although the current explanation makes intuitive sense, this approach does not seem to be used in previous work. The authors could design a simulation that supports their approach.*

Reply: We felt that weighted PCA was a nice but unnecessary complication to this work, so have removed it entirely.

(2.4) *It is conceivable that some subpopulations will have more missingness in some windows. That may skew the resulting PCs by selecting different sample sizes for the different windows (as discussed in Appendix B). This could distort the PCs, so that variation reflects underlying variation in missingness. Would be good to discuss this potential issue and provide simulations.*

Reply: We have tested the method against a large quantity of missing data whereby the amount of missingness does not vary by population (p. 31, l. 821). However, missingness that is structured by population could absolutely cause variation in PCs. This could happen, for instance, with RAD data wherein allelic dropout is driven by population-specific polymorphism in the restriction site. The same problem to a lesser degree would be caused by mapping bias, and in fact we see windows around the breakpoints of the polymorphic inversions of the DPGP data having higher rates of missingness. We definately would like our method to be insensitive to missing data, but if missingness is in fact correlated with changes in patterns of relatedness (as in these cases), correlation of MDS scores with amount of missingness is not in fact a problem. As we don't think this point

is particularly helpful to the reader, we have provided a plot of missingness against MDS score (Figure S18) and mentioned in the text (p. 10, l. 282), but have not added a lengthy discussion.

(2.5) Appendix A: when using jackknife to estimate variance, each window is being divided in 10 “independent” resampling units. Due to LD, these 10 blocks are likely correlated, which would bias the estimates of variance. This is probably not a problem because both signal and noise could be equally biased, but the authors may want to consider this potential issue. I wonder if the correlation with recombination rate may be partially explained by this.

Reply: That’s true, but we are unaware of a better option, and since the jackknife is only used to pick an appropriate window size, it does not seem crucial. (p. 29, l. 788) Since results are the same across different window sizes, we don’t think this can explain a correlation with recombination rate.

(2.6) Is it possible to explain the results of Figure 6 just considering neutral variation in local ancestry due to recent admixture? This may explain why ancestry seems to explain a fair amount of variance in the lower plots of Fig 6. Local PCA has been previously used by others to detect local ancestry blocks, e.g. see the PCAdmix approach by Brisbin et al. The authors discuss the possibility that admixture is driving the differentiation, but do not test whether their observations agree with neutrality.

Reply: This is a good point, but any *neutral* variation in local ancestry is not expected to vary systematically along the chromosome, as we now discuss more explicitly (p. 16, l. 394). It well could be that the patterns we see are due to selection acting differentially on introgressed segments, a scenario that would fall under “linked selection”. We discuss this in the Discussion (p. 19, l. 406). We also now cite and discuss the differences to PCAdmix (p. 3, l. 79).

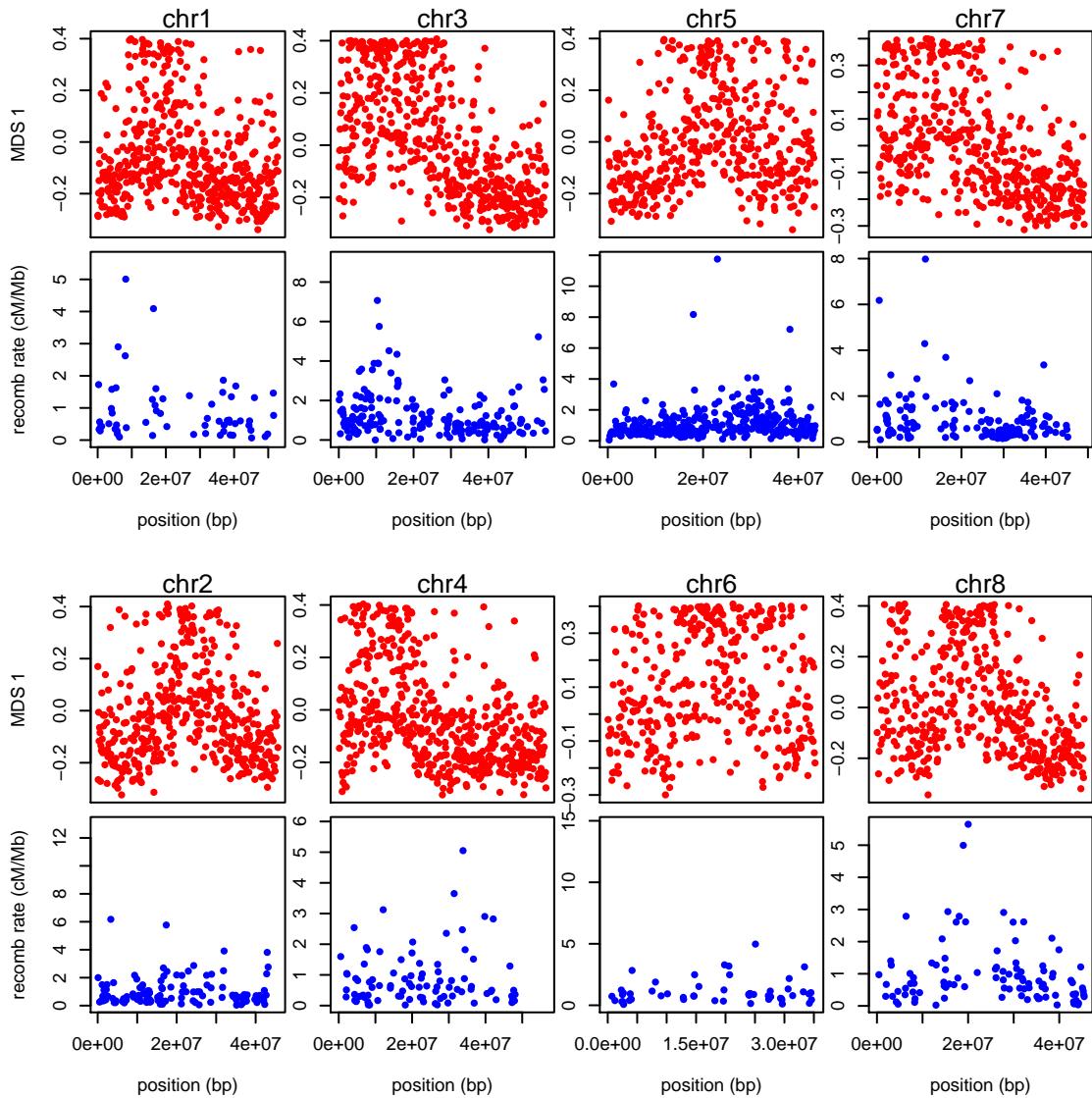
(2.7) “to remove the effect of artifacts such as mutation rate variation, we also rescale each approximate covariance matrix to be of similar size (precisely, so that the underlying data matrix has trace norm equal to one”. This potential issue is a bit unclear to me, since I would expect that scaling the volume of local trees would not result in changed distances in PC space. Perhaps the authors could show via simulations that this creates a problem, and that the normalization addresses it.

Reply: We agree that demonstrating this issue would be nice in principle, but for this audience would be getting too far down in the weeds, and including the requested simulations would inflate an already lengthy paper. Additionally, we aim to provide a good method, not to prove that the method is better than all other possible methods. For the reviewer’s sake, here is some more about our rationale. Suppose that the covariance matrix is the same along the whole genome, except that it is multiplied by an overall constant

that varies. Since PCs are scaled to have norm 1, this overall scaling will indeed not affect the PCs, but we don't compare the PCs directly – we compare the low-dimensional approximations to the covariance matrix, and the scaling we propose would remove the effect of this overall constant. It is not clear that variation in mutation rate would act like an overall multiple of the covariance matrix, it since we're interested in patterns rather than magnitudes of relatedness, it seems a wise thing to remove.

(2.8) Figure 7: Are MDS coordinates correlated with recombination rates in this case?

Reply: We made a stab at checking this, and obtained the best version to date of the *Medicago* recombination map from Tim Paape and Peter Tiffin. There are two versions: a very coarse physical map, and a fine-scale map estimated using LDhat. However, both are on version 3.0 of the assembly, while all other coordinates (sequencing data; gene annotations) are in version 4.0. Furthermore, as Peter Tiffin told us, “apparently there are no files that translate Mt3.0 to Mt4.0 locations (yes, seems a bit silly).” There is a liftOver chain file for translating 3.5 to 4.0, and “the differences in the Mt3.0 and Mt3.5 assemblies are, however, apparently relatively minor”. On this basis, we produced the desired figure assuming that Mt3.0 coordinates are the same as Mt3.5 coordinates, included to satisfy the reviewers’ curiosity:



However, given uncertainties in this mapping, the relatively poor match of window sizes, the large number of unmappable windows, and the nature of the recombination data (produced with LDhat, not with actual observations of recombinations), we decided not to include this (but have provided a note, (p. 14, l. 363)).

(2.9) Application: *Is what the authors seem to be proposing not already accounted for by linear mixed model association approaches? If not, this should be clarified. Either way, this paragraph could be dropped.*

Reply: It is not accounted for; we've added some more explanation to this section (which

we think should stay in). (p. 20, l. 448)

(2.10) Introduction: “it is not necessarily clear what aspects of demography should be included in the concept.” I find it a bit weird to describe selection as an “aspect of demography”. Although it could be seen as such within a coalescent framework, that seems to be just a useful representation. The authors may consider rewording.

Reply: Good point, but we think that the sense of unease is perhaps desireable here: certainly differential survival is an aspect of demography; but as the rest of the paragraph goes on to discuss, this quickly spills over into selection, which probably should not be included. We've opted to keep this. (p. 2, l. 37)

(2.11) Paragraph starting in “Since the definition...”. The notation is a bit unclear. Please check that it is clear which PC the text refers to.

Reply: We've rewritten this section; hopefully it is clearer. (p. 29, l. 779)

(2.12) Would the authors be able to provide a sense for the directionality of effects in Figure 4? It would be interesting if the authors tried to further characterize regions that are similar due to higher recombination rates. E.g. is there more/less density of polymorphisms in these regions?

Reply: We agree, that further investigation into these patterns and their correlates would be very interesting and fruitful; however, we think it's outside of the scope of this paper.

(2.13) Page 13: typo: “figures 6 and 6”.

Reply: Fixed. (p. 14, l. 352)

(2.14) Typo in abstract, line 6 “, We show” → “. We show”.

Reply: Fixed.

(2.15) Typo: end of introduction “an visualization”. The whole sentence is a bit weird. The authors just stated focus is on clustering, not on looking for outliers, but what does it mean that “we allow ourselves to be surprised by unexpected signals in the data”?

Reply: We have removed this sentence.

(2.16) “There has been substantial debate over the relative impacts of different forms of selection.” Citation needed.

Reply: We have added some relevant recent citations, but apologize for not being aware of the appropriate recent review to refer to. (p. 19, l. 414)

(2.17) “Results using larger numbers of PCs were nearly identical”. It would be interesting to have a supplementary table.

Reply: We have included a table of correlations (Table S2), as well as the plots for Drosophila data using $k = 5$ (Figure S2), and hope that readers will take our word on this point more generally.

(2.18) Table 1 legend seems a bit redundant. Columns are self-explanatory.

Reply: Good point; we've cut this down.

(2.19) It would help to have numbered lines and references.

Reply: We've made these changes.