

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 (1949) defined *population structure* to encompass “such matters as numbers, com-
33 position 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 genetic
35 patterns that result from this process, as for instance reduced mean relatedness between
36 individuals from distinct populations. However, it is not necessarily clear what aspects of
37 demography should be included in the concept. For instance, Blair (1943) defines *popula-*
38 *tion structure* to be the sum total of “such factors as size of breeding populations, periodic
39 fluctuation of population size, sex ratio, activity range and *differential survival of progeny*”
40 (emphasis added). The definition is similar to Wright’s, but differs in including the effects
41 of natural selection. On closer examination, incorporating differential survival or fecundity
42 makes the concept less clear: should a randomly mating population consisting of two types
43 that are partially reproductively isolated from each other be said to show population struc-
44 ture or not? Whatever the definition, it is clear that due to natural selection, the effects
45 of population structure – the *realized* patterns of genetic relatedness – differ depending on
46 which portion of the genome is being considered. For instance, strongly locally adapted
47 alleles of a gene will be selected against in migrants to different habitats, increasing genetic
48 differentiation between populations near to this gene. Similarly, newly adaptive alleles
49 spread first in local populations. These observations motivate many methods to search for
50 genetic loci under selection, as for example in Huerta-Sánchez et al. (2013), Martin et al.
51 (2016), and Duforet-Frebbourg et al. (2015).

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 (Barton 2000; Charlesworth et al. 1993; Kim and Stephan 2002). Indeed, many
59 species show chromosome-scale variation in diversity and divergence (e.g., (Langley et al.
60 2012)); species phylogenies can differ along the genome due to incomplete lineage sorting,
61 adaptive introgression and/or local adaptation (e.g., Ellegren et al. (2012); Nadeau et al.
62 (2012); Pease and Hahn (2013); Pool (2015); Vernot and Akey (2014)); and theoretical
63 expectations predict that geographic patterns of relatedness should depend on selection
64 (Charlesworth et al. 2003).

65 Patterns in genome-wide relatedness are often summarized by applying principal com-
66 ponents analysis (PCA, Patterson et al. (2006)) to the genetic covariance matrix, as pio-
67 neered by Menozzi et al. (1978). The results of PCA can be related to the genealogical
68 history of the samples, such as time to most recent common ancestor and migration rate
69 between populations (McVean 2009; Novembre and Stephens 2008), and sometimes pro-
70 duce “maps” of population structure that reflect the samples’ geographic origin distorted

71 by rates of gene flow (Novembre et al. 2008).

72 Modeling such “background” kinship between samples is essential to genome-wide association studies (GWAS, Astle and Balding (2009); Price et al. (2006)), and so understanding variation in kinship along the genome could lead to more generally powerful methods, and may be essential for doing GWAS in species with substantial heterogeneity in realized patterns of mean relatedness along the genome.

77 PCA has been applied to genomic windows in methods to infer tracts of local ancestry
78 in recently admixed populations (Brisbin et al. 2012; Bryc et al. 2010), and to identify
79 putative chromosomal inversions (Ma and Amos 2012).

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 (1997); Manjón et al. (2013); Roweis and Saul (2000); Weingessel and Hornik
84 (2000)), performing PCA not on the entire dataset but instead on subsets of observations,
85 providing local pictures which are then stitched back together to give a global picture. At
86 first 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. (2012); Turner
93 et al. (2005)). However, a given pattern (e.g., valleys of F_{ST}) can be caused by more than
94 one biological process (Burri et al. 2015; Cruickshank and Hahn 2014), which in retrospect
95 is unsurprising given that we are using a single statistic to describe a complex process.
96 It is also common to use methods such as PCA to visualize large-scale patterns in mean
97 genome-wide relatedness. In this paper we show if and how patterns of mean relatedness
98 vary systematically along the genome, in a way particularly suited to large samples from
99 geographically distributed populations. Geographic population structure sets the stage by
100 establishing “background” patterns of relatedness; our method then describes how this
101 structure is affected by selection and other factors. Our aim is not to identify outlier
102 loci, but rather to describe larger-scale variation shared by many parts of the genome;
103 correlation of this variation with known genomic features can then be used to uncover its
104 source.

105 2 Materials and Methods

106 As depicted in Figure 1, the general steps to the method are: (1) divide the genome into
107 windows, (2) summarize the patterns of relatedness in each window, (3) measure dissimilarity
108 in relatedness between each pair of windows, (4) visualize the resulting dissimilarity

109 matrix using multidimensional scaling (MDS), and (5) combine similar windows to more
110 accurately visualize local effects of population structure using PCA.

111 2.1 PCA in genomic windows

112 To begin, we first recoded sampled genotypes as numeric matrices in the usual manner, by
113 recording the number of nonreference alleles seen at each locus for each sample. We then
114 divided the genome into contiguous segments (“windows”) and applied principal compo-
115 nent analysis (PCA) as described in McVean (2009) separately to the submatrices that
116 corresponded to each window. The choice of window length entails a tradeoff between sig-
117 nal and noise, since shorter windows allow better resolution along the genome but provide
118 less precise estimates of relatedness. A method for choosing a window length to balance
119 these considerations is given in Appendix A. Precisely, denote by Z the $L \times N$ recoded
120 genotype matrix for a given window (L is the number of SNPs and N is the sample size),
121 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
122 the sum is over the n_s nonmissing genotypes. We first compute the mean-centered ma-
123 trix X , as $X_{si} = Z_{si} - \bar{Z}_s$, and preserving missingness. (This mean-centering makes the
124 result not depend on the choice of reference allele, exactly if there is no missing data,
125 and approximately otherwise.) Next, we find the covariance matrix of X , denoted C , as
126 $C_{ij} = \frac{1}{m_{ij}-1} \sum_s X_{si} X_{sj} - \frac{1}{m_{ij}(m_{ij}-1)} (\sum_s X_{si})(\sum_s X_{sj})$, where all sums are over the m_{ij}
127 sites where both sample i and sample j have nonmissing genotypes. The principal com-
128 ponents are the eigenvectors of C , normalized to have Euclidean length equal to one, and
129 ordered by magnitude of the eigenvalues.

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

136 Several of the datasets we use have unbalanced representations of diverged populations,
137 which can have a strong impact on the results of PCA. (The principal axes may describe
138 variation *within* an overrepresented group rather than more significant variation between
139 groups.) Therefore, to check that sampling patterns do not affect our results, we compared
140 to a variant of PCA that gives roughly equal weight to each group of samples, rather
141 than to each sample. The rationale and implementation of this method are described in
142 Appendix B.

143 2.2 Similarity of patterns of relatedness between windows

144 We think of the local effects of population structure as being summarized by *relative*
145 position of the samples in the space defined by the top principal components. However, we

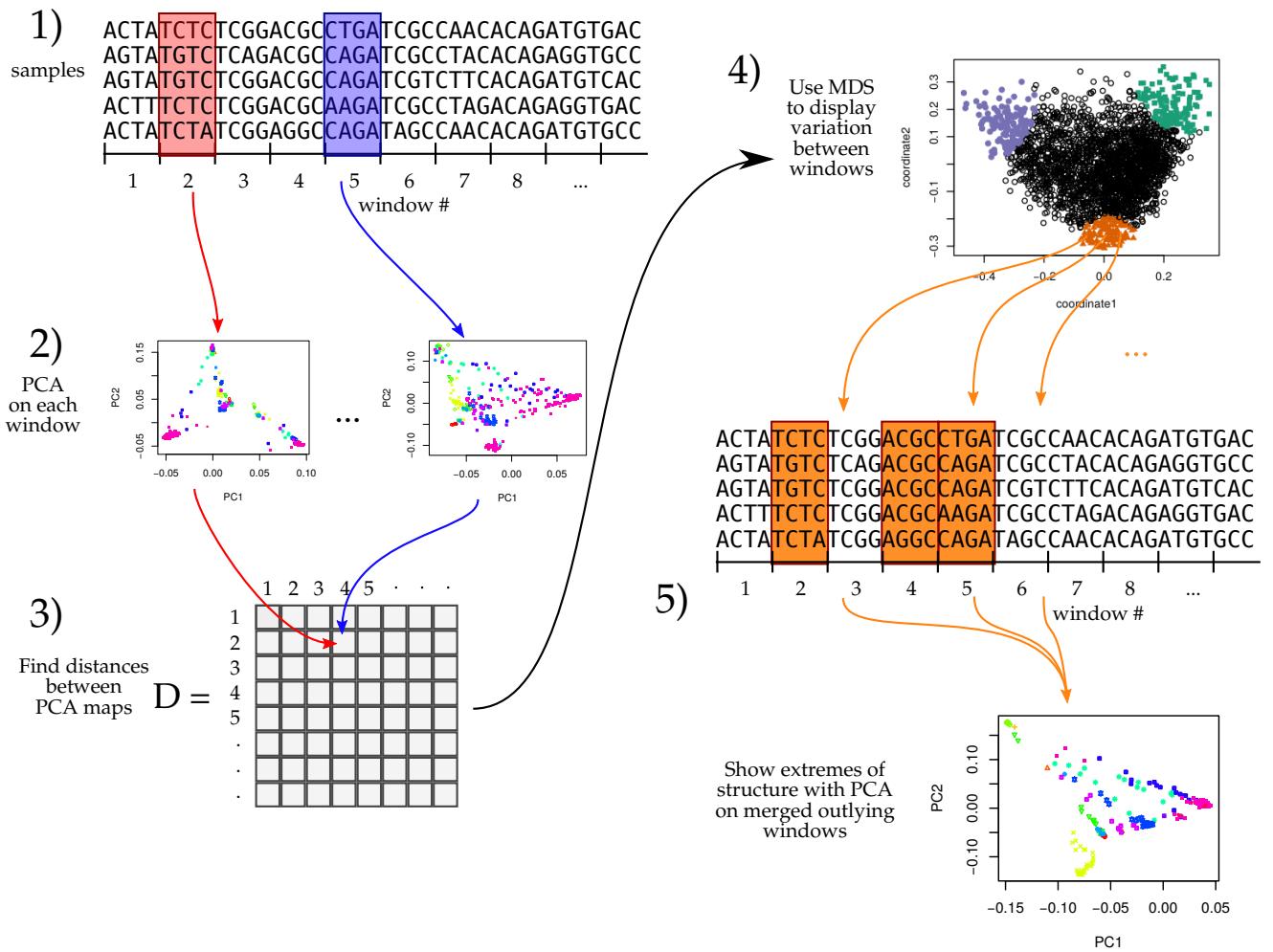


Figure 1: An illustration of the method; see Methods for details.

146 do not compare patterns of relatedness of different genomic regions by directly comparing
 147 the PCs, since rotations or reflections of these imply identical patterns of relatedness.
 148 Instead, we compare the low-dimensional approximations of the local covariance matrices
 149 obtained using the top k PCs, which is invariant under ordering of the PCs, reflections, and
 150 rotations and yet contains all other information about the PCs. (For results shown here,
 151 we use $k = 2$; results using larger numbers of PCs were nearly identical.) Furthermore,
 152 to remove the effect of artifacts such as mutation rate variation, we also rescale each
 153 approximate covariance matrix to be of similar size (precisely, so that the underlying data
 154 matrix has trace norm equal to one).

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

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

159 To measure the similarity of patterns of relatedness for the i^{th} window and j^{th} win-
 160 dows, we then use Euclidean distance D_{ij} between the matrices $M(i)$ and $M(j)$: $D_{ij}^2 =$
 161 $\sum_{k\ell} (M(i)_{k,\ell} - M(j)_{k,\ell})^2$.

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

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

170 2.3 Visualization of results

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

176 We then locate variation in patterns of relatedness along the genome by choosing col-
 177 lections of windows that are nearby in MDS coordinates, and map their positions along the

178 genome. A visualization of the effects of population structure across the entire collection
179 is formed by extracting the corresponding genomic regions and performing PCA on all,
180 aggregated, regions.

181 2.4 Testing

182 We tested the method using two types of simulation. First, to verify expected behavior,
183 we simulated “genomes” as an independent sequence of correlated Gaussian “genotypes”,
184 using a different covariance matrix in the first quarter, middle half, and last quarter of
185 the chromosome. To verify robustness to missing data, we ran the method after randomly
186 dropping 50% of the genotypes in the first half of the genome; if the method is misled by
187 missing data, then it will distinguish the two halves of the chromosome rather than the
188 segments having different covariance matrices. Details are given in Appendix ??.

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

208 2.5 Datasets

209 We applied the method to genomic datasets with good geographic sampling: 380 African
210 *Drosophila melanogaster* from the Drosophila Genome Nexus (Lack et al. 2015), a world-
211 wide dataset of humans, 3,965 humans from several locations worldwide from the POPRES
212 dataset (Nelson et al. 2008), and 263 *Medicago truncatula* from 24 countries around the
213 Mediterranean basin a range-wide dataset of the partially selfing weedy annual plant from
214 the *Medicago truncatula* Hapmap Project (Tang et al. 2014), as summarized in Table 1.

215 ***Drosophila melanogaster*:** We used whole-genome sequencing data from the Drosophila
216 Genome Nexus (<http://www.johnpool.net/genomes.html>, (Lack et al. 2015)), consisting
217 of the Drosophila Population Genomics Project phases 1–3 (Langley et al. 2012; Pool
218 et al. 2012), and additional African genomes (Lack et al. 2015). After removing 20 genomes
219 with more than 8% missing data, we were left with 380 samples from 16 countries across
220 Africa and Europe. Since the *Drosophila* samples are from inbred lines or haploid embryos,
221 we treat the samples as haploid when recoding; regions with residual heterozygosity were
222 marked as missing in the original dataset; we also removed positions with more than 20%
223 missing data. Each chromosome arm we investigated (X, 2L, 2R, 3L, and 3R) has 2–3
224 million SNPs; PCA plots for each arm are shown in Figure S2.

225 **Human:** We also used genomic data from the entire POPRES dataset (Nelson et al.
226 2008), which has array-derived genotype information for 447,267 SNPs across the 22 au-
227 tosomes of 3,965 samples in total: 346 African-Americans, 73 Asians, 3,187 Europeans
228 and 359 Indian Asians. Since these data derive from genotyping arrays, the SNP density is
229 much lower than the other datasets, which are each derived from whole genome sequencing.
230 We excluded the sex chromosomes and the mitochondria. PCA plots for each chromosome,
231 separately, are shown in Figure S3.

232 ***Medicago truncatula*:** Finally, we used whole-genome sequencing data from the *Med-*
233 *icago truncatula* Hapmap Project (Tang et al. 2014), which has 263 samples from 24 coun-
234 tries, primarily distributed around the Mediterranean basin. Each of the 8 chromosomes
235 has 3–5 million SNPs; PCA plots for these are shown in Figure S4. We did not use the
236 mitochondria or 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.

237 2.6 Data access

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

241 Datasets are available as follows: human (POPRES) at dbGaP with accession number
242 phs000145.v4.p2, *Medicago* at the Medicago Hapmap <http://www.medicagohapmap.org/>,
243 and *Drosophila* at the Drosophila Genome Nexus, <http://www.johnpool.net/genomes.html>.
244

245 3 Results

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

252 3.1 Validation

253 *Address: mutation rate variation; recombination rate variation; choice of number of PCs;*
254 *choice of window size; variation in missingness; maybe differing sample sizes and reweighting.*

255 Simple non-population-based simulations with Gaussian “genotypes” showed that the
256 method performs as expected, clearly separating regions of the genome with different underly-
257 ing covariance matrices without being affected by extreme differences in amount of
258 missing data (Supplemental Figure S24). This simulation also verifies insensitivity to or-
259 dering of top PCs, since it was performed using a covariance matrix with the top two
260 eigenvalues equal, so that the order of empirical eigenvectors (PCs) switches randomly.
261 The same type of simulations with dramatically different sample sizes showed that using
262 weighted PCA substantially improved identification of different genomic regions (Supple-
263 mental Figure S25).

264 Individual-based simulations using SLiM (?) allowed us to test the effects of recom-
265 bination and mutation rate variation, as well as linked selection. Results are shown in
266 Supplemental Figures S15 and ???. As expected, varying recombination rate stepwise by
267 a factor of 64 did not induce patterns in the MDS visualizations correlated with recom-
268 bination rate. Since varying mutation rate with a fixed recombination map is equivalent
269 to varying the recombination map and remapping windows, we used the same simulations
270 to show that variation in mutation rate similarly has no effect. On the other hand, a
271 recombination map with hotspots (the HapMap human female map for chromosome 7 (?))
272 induced outliers at long regions of low recombination rate (also as expected).

273 Simulations with linked selection produced mixed results (Figure ??). The method
274 applied to spatially varying linked selection strongly identified the regions under selection,
275 but did not clearly distinguish those regions with weaker levels of constant selection. These

277 tests are not meant to be comprehensive survey of linked selection, but only to demonstrate
278 that linked selection can produce signals similar to what we see in real data.

279 **3.2 *Drosophila melanogaster***

280 We applied the method to windows of average length 9 Kbp, across chromosome arms
281 2L, 2R, 3L, 3R and X separately. The first column of Figure 2 is a multidimensional
282 scaling (MDS) visualization of the matrix of dissimilarities between genomic windows: in
283 other words, genomic windows that are closer to each other in the MDS plot show more
284 similar patterns of relatedness. For each chromosome arm, the MDS visualization roughly
285 resembles a triangle, sometimes with additional points. Since the relative position of each
286 window in this plot shows the similarity between windows, this suggests that there are
287 at least three extreme manifestations of population structure typified by windows found
288 in the “corners” of the figure, and that other windows’ patterns of relatedness may be a
289 mixture of those extremes. The next two columns of Figure 2 respectively depict the two
290 MDS coordinates of each window, plotted against the window’s position along the genome,
291 to show how the plot of the first column is laid out along the genome.

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

300 The most striking variation in patterns of relatedness turns out to be explained by
301 several large inversions that are polymorphic in these samples, discussed in Corbett-Detig
302 and Hartl (2012) and Langley et al. (2012). To depict this, Figure 3 shows the PCA plots in
303 Figure S1 recolored by the orientation of the inversion for each sample. Taking chromosome
304 arm 2L as an example, the two regions of similar, extreme patterns of relatedness shown
305 in green in the first row of Figure 2 lie directly around the breakpoints of the inversion
306 In(2L)t, and the PCA plots in the first rows of Figure 3 shows that patterns of relatedness
307 here are mostly determined by inversion orientation. The regions shown in purple on
308 chromosome 2L lie near the centromere, and have patterns of relatedness reflective of two
309 axes of variation, seen in Figures S1 and 3, which correspond roughly to latitude within
310 Africa and to degree of cosmopolitan admixture respectively (see Lack et al. (2015) for more
311 about admixture in this sample). The regions shown in orange on chromosome 2L mostly lie
312 inside the inversion, and show patterns of relatedness that are a mixture between the other
313 two, as expected due to recombination within the (long) inversion (Guerrero et al. 2011).
314 Similar results are found in other chromosome arms, albeit complicated by the coexistence
315 of more than one polymorphic inversion; however, each breakpoint visibly affects patterns

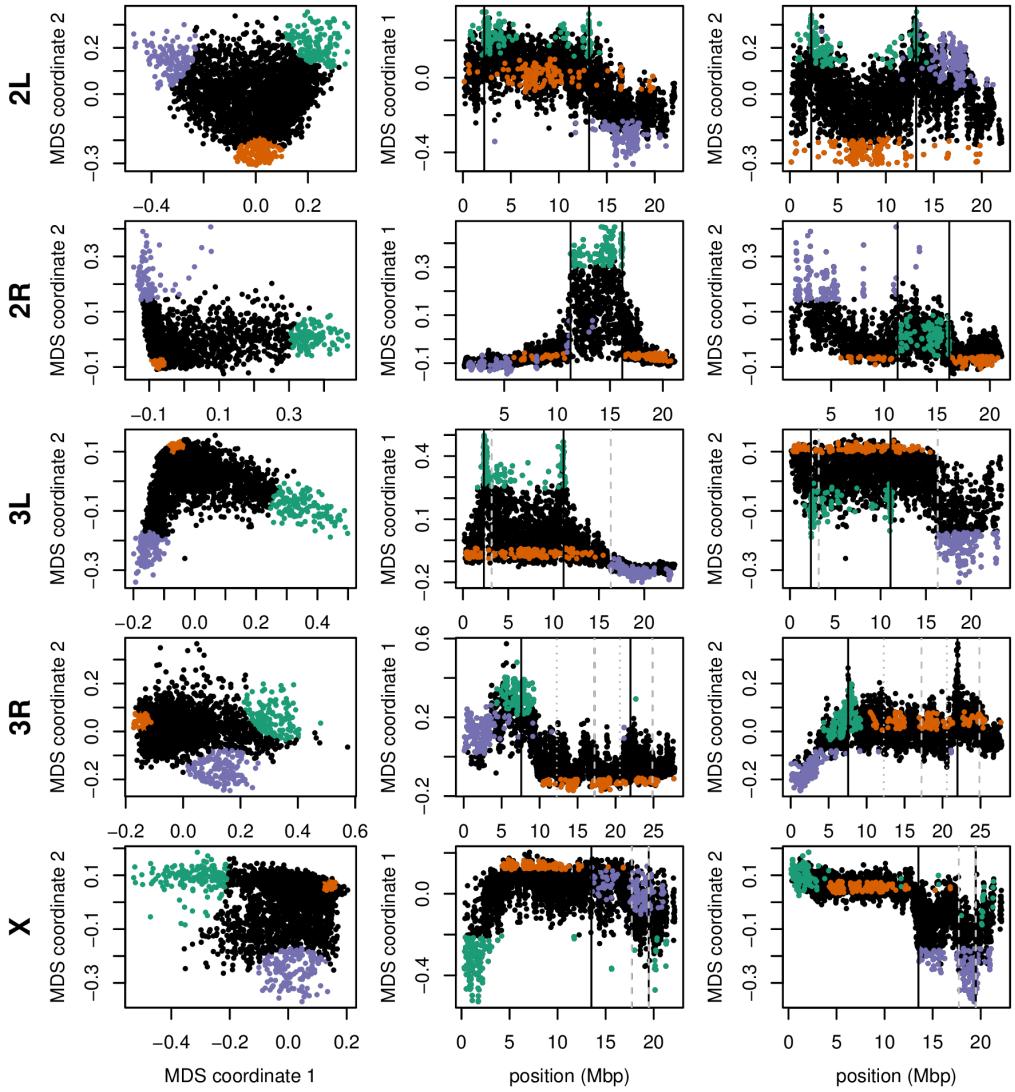


Figure 2: Variation in patterns of relatedness for windows across *Drosophila melanogaster* chromosome arms. In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the two MDS 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.

316 in the MDS coordinates (see vertical lines in Figure 2).

317 To see how patterns of relatedness vary in the absence of polymorphic inversions, we
318 performed the same analyses after removing, for each chromosome arm, any samples car-
319 rying inversions on that arm. In the result, shown in Supplemental Figure S5, the striking
320 peaks associated with inversion breakpoints are gone, and previously smaller-scale vari-
321 ation now dominates the MDS visualization. For instance, the majority of the variation
322 along 3L in Figure 2 is on the left end of the arm, dominated by two large peaks around the
323 inversion breakpoints; there is also a relatively small dip on the right end of the arm (near
324 the centromere). In contrast, Supplemental Figure S5 shows that after removing polymor-
325 phic inversions, remaining structure is dominated by the dip near the centromere. Without
326 inversions, variation in patterns of relatedness shown in the MDS plots follows similar pat-
327 terns to that previously seen in *D. melanogaster* recombination rate and diversity (Langley
328 et al. 2012; Mackay et al. 2012). Indeed, correlations between the recombination rate in
329 each window and the position on the first MDS coordinate are highly significant (Spear-
330 man’s $\rho = 0.54$, $p < 2 \times 10^{-16}$; Figures 4 and S6). This is consistent with the hypothesis
331 that variation is due to selection, since the strength of linked selection increases with local
332 gene density, measured in units of recombination distance. The number of genes – mea-
333 sured as the number of transcription start and end sites within each window – was not
334 significantly correlated with MDS coordinate ($p = 0.22$).

335 3.3 Human

336 As we did for the *Drosophila* data, we applied our method separately to all 22 human
337 autosomes. On each, variation in patterns of relatedness was dominated by a small number
338 of windows having similar patterns of relatedness to each other that differed dramatically
339 from the rest of the chromosome. These may be primarily inversions: outlying windows
340 coincide with three of the six large polymorphic inversions described in Antonacci et al.
341 (2009), notably a particularly large, polymorphic inversion on 8p23 (Figure 5). Similar
342 plots for all chromosomes are shown in Supplementary Figures S7, S8, and S9. PCA plots
343 of many outlying windows show a characteristic trimodal shape (shown for chromosome 8 in
344 Figure S10), presumably distinguishing samples having each of the three diploid genotypes
345 for each inversion orientation (although we do not have data on orientation status). This
346 trimodal shape has been proposed as a method to identify inversions (Ma and Amos 2012),
347 but distinguishing this hypothesis from others, such as regions of low recombination rate,
348 would require additional data.

349 We also applied the method on all 22 autosomes together, and found that, remarkably,
350 the inversion on chromosome 8 is still the most striking outlying signal (Figure S11).
351 Further investigation with a denser set of SNPs, allowing a finer genomic resolution, may
352 yield other patterns.

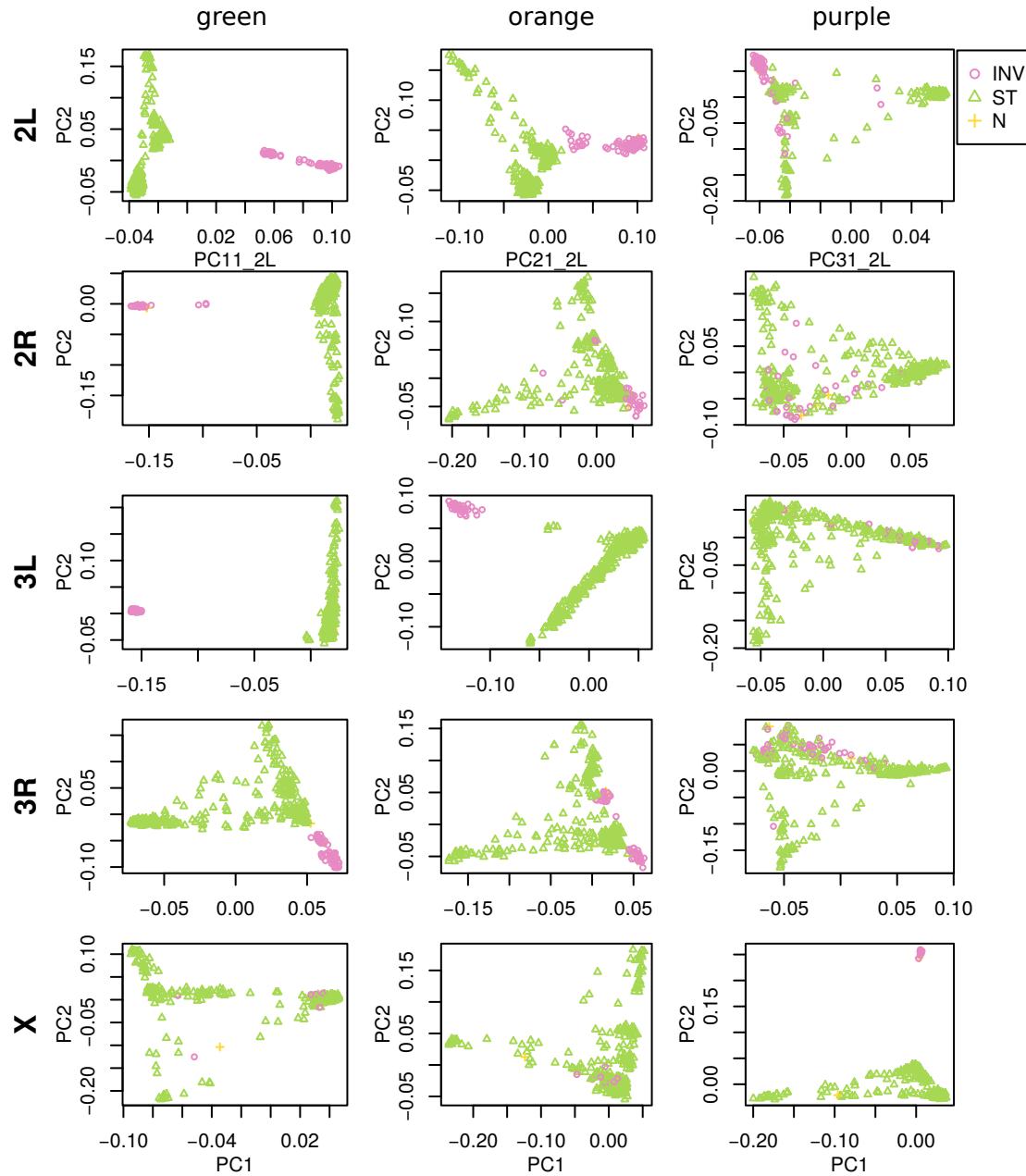


Figure 3: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange 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 (Lack et al. 2015)). In each “INV” denotes an inverted genotype, “ST” denotes the standard orientation, and “N” denotes unknown.

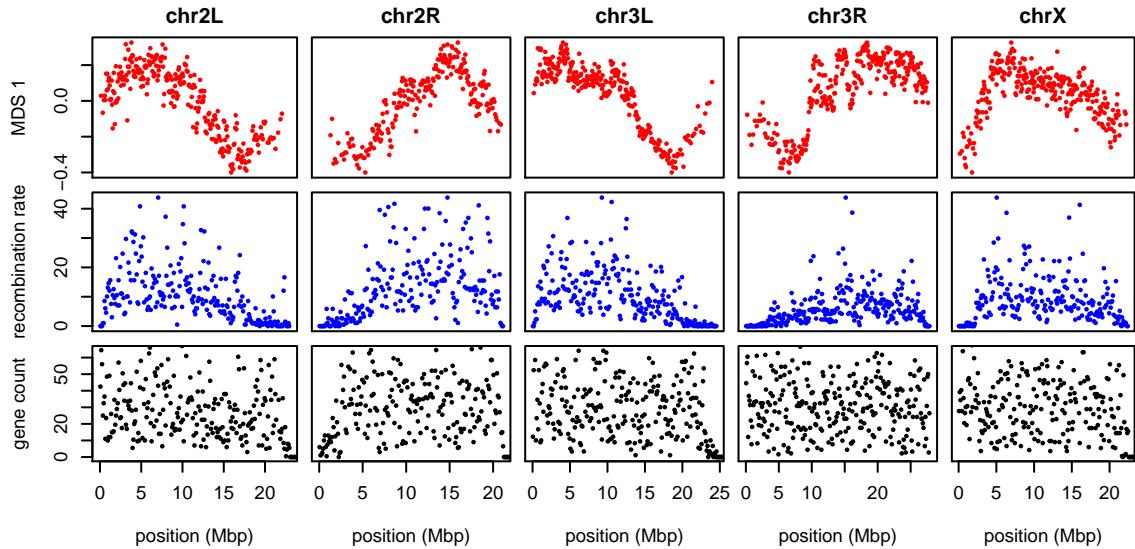


Figure 4: The effects of population structure without inversions is correlated to recombination rate in *Drosophila melanogaster*. The first plot (in red) shows the first MDS coordinate along the genome for windows of 10,000 SNPs, obtained after removing samples with inversions. (A plot analogous to Figure 2 is shown in Supplemental Figure S5.) 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 (Fiston-Lavier et al. 2010) release 5, using rates from Comeron et al. (2012). 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.

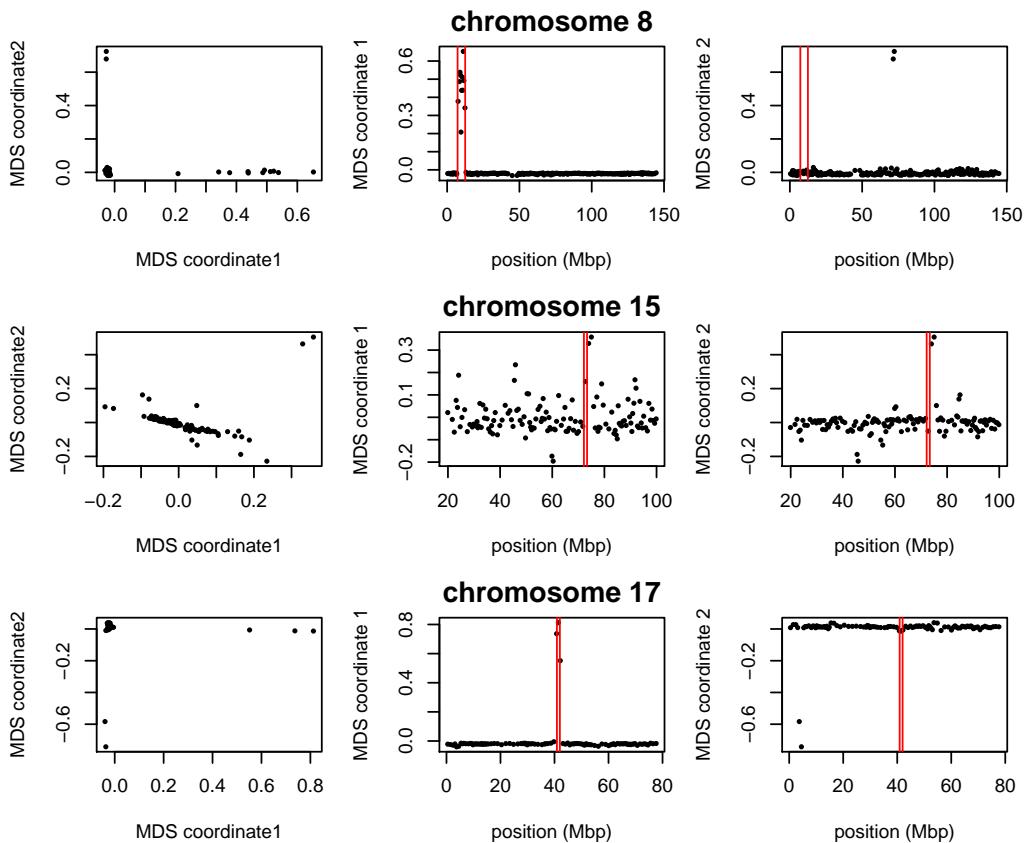


Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in each plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position (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. (2009).

353 **3.4 *Medicago truncatula***

354 Unlike the other two species, the method applied separately on all eight chromosomes of
355 *Medicago truncatula* showed similar patterns of gradual change in patterns of relatedness
356 across each chromosome, with no indications of chromosome-specific patterns. This con-
357 sistency suggests that the factor affecting the population structure for each chromosome
358 is the same, as might be caused by varying strengths of linked selection. To verify that
359 variation in the effects of population structure is shared across chromosomes, we applied
360 the method to all chromosomes together. Results for chromosome 3 are shown in Figures
361 6 and 6, and other chromosomes are similar: across chromosomes, the high values of the
362 first MDS coordinate coincide with the position of the heterochromatic regions surrounding
363 the centromere, which often have lower gene density and may therefore be less subject to
364 linked selection. To verify that this is a possible explanation, we counted the number of
365 genes found in each window using gene models in Mt4.0 from jcvi.org (Tang et al. 2014),
366 which are shown juxtaposed with the first MDS coordinate of each window in Figure 7,
367 and are significantly correlated, as shown in Supplemental Figure S12. (Values shown are
368 the number of start and end positions of each predicted mRNA transcript, divided by two,
369 assigned to the nearest window.) However, other genomic features, such as distance to
370 centromere show roughly the same patterns, so we cannot rule out alternative hypotheses.
371 In particular, fine-scale recombination rate estimates are not available in a form mappable
372 to Mt4.0 coordinates (although those in Paape et al. (2012) appear visually similar).

373 We also found nearly identical results when choosing shorter windows of 1,000 SNPs;
374 or choosing windows of equal length in base pairs rather than SNPs. Similarly, the results
375 were not substantially changed when using weighted PCA to downweight the large group
376 of Tunisian samples.

377 **4 Discussion**

378 Our investigations have found substantial variation in the patterns of relatedness formed
379 by population structure across the genomes of three diverse species, revealing distinct bi-
380 logical processes driving this variation in each species. More investigation, particularly
381 on more species and datasets, will help to uncover what aspects of species history can
382 explain these differences. With growing appreciation of the heterogeneous effects of se-
383 lection across the genome, especially the importance of adaptive introgression and hybrid
384 speciation (Brandvain et al. 2014; Fitzpatrick et al. 2010; Hufford et al. 2013; Pool 2015;
385 Staubach et al. 2012), local adaptation (Lenormand 2002; Wang and Bradburd 2014), and
386 inversion polymorphisms (Kirkpatrick 2010; Kirkpatrick and Barrett 2015), local PCA may
387 prove to be a useful exploratory tool to discover important genomic features.

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

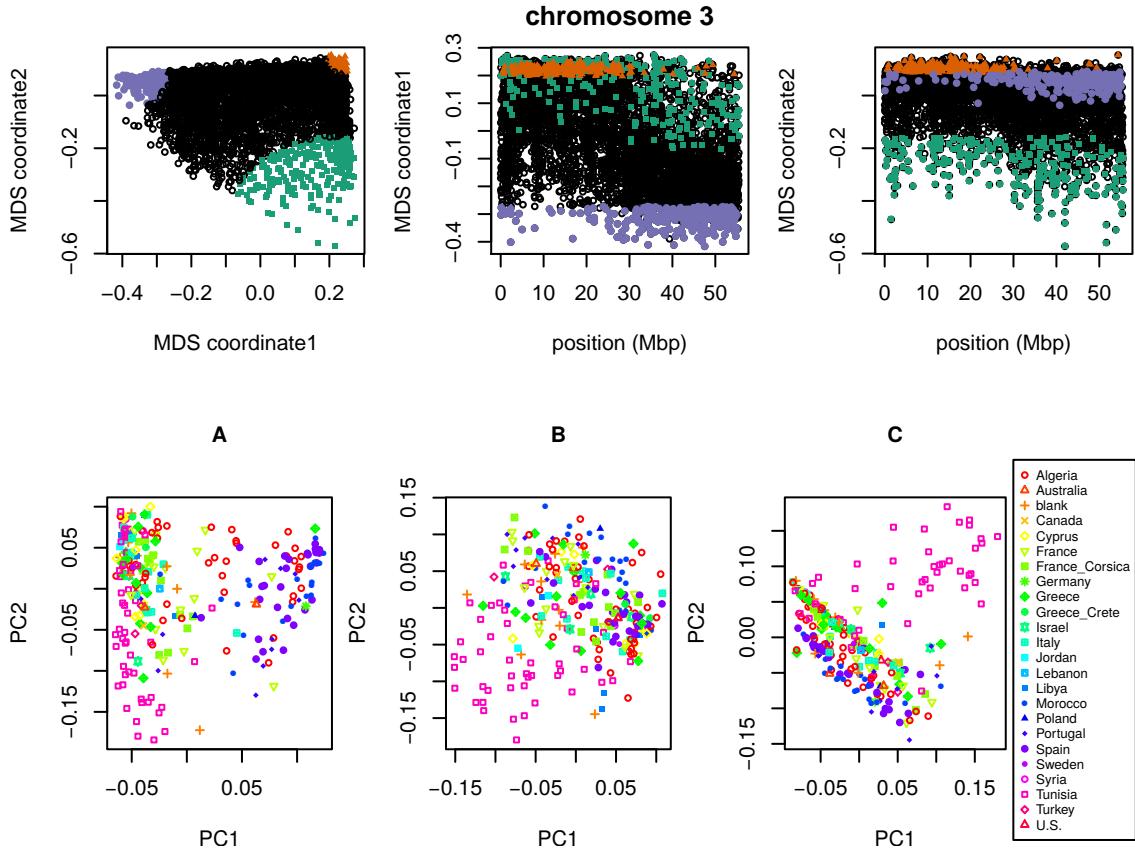


Figure 6: MDS visualization of patterns of relatedness on *M. truncatula* chromosome 3, with corresponding PCA plots. Each point in the plot represents a window; the structure revealed by the MDS plot is strongly clustered along the chromosome, with windows in the upper-right corner of the MDS plot (colored 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 S13. **(below)** PCA plots for the sets of genomic windows colored (A) green, (B) orange, and (C) purple in Figure 6. Each point corresponds to a sample, colored by country of origin. Plots for remaining chromosomes are shown in Supplemental Figure S14.

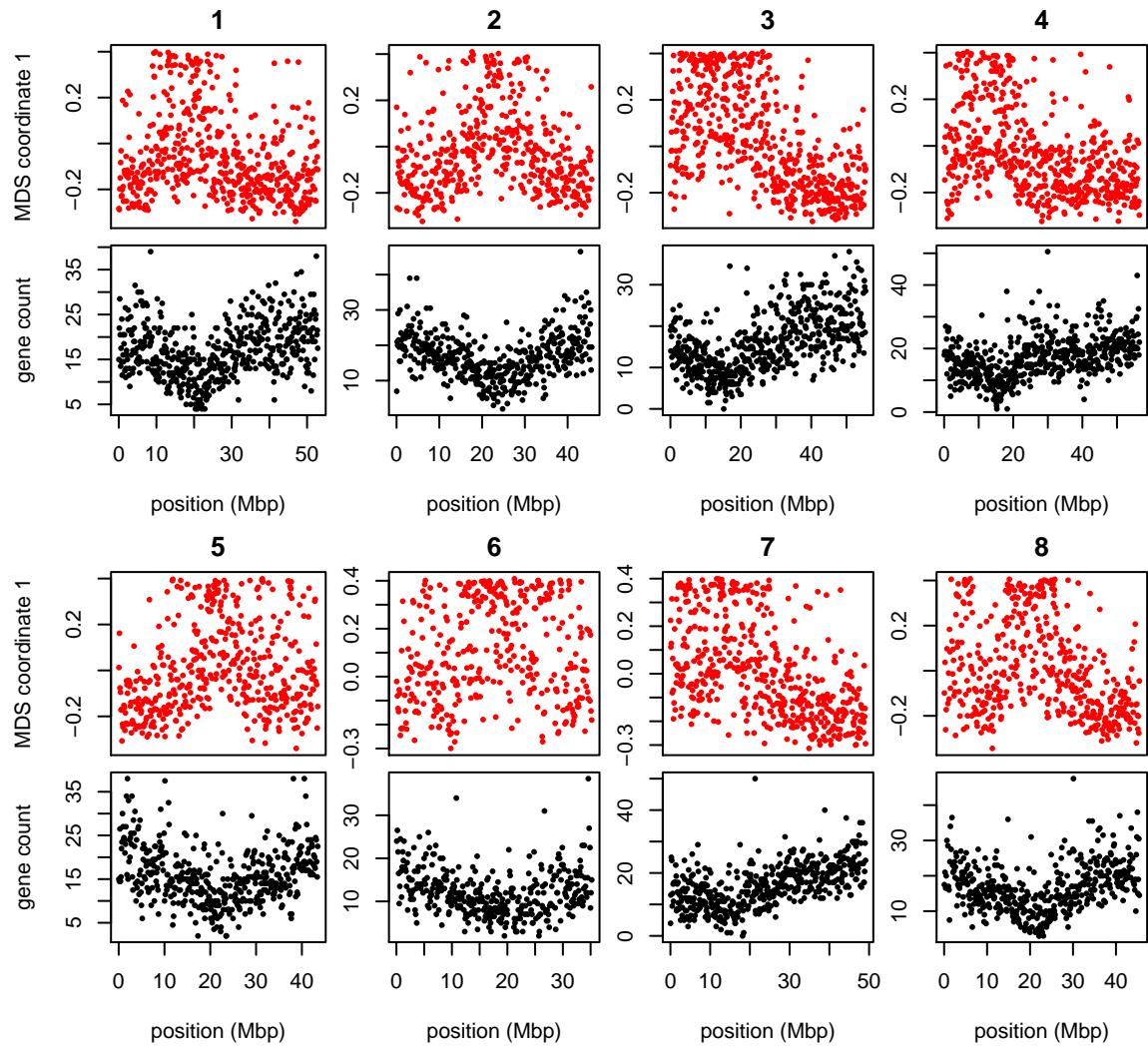


Figure 7: MDS coordinate and gene density for each window in the *Medicago* genome, for chromosomes 1–8 (numbered above each pair of figures). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene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

391 **Chromosomal inversions** A major driver of variation in patterns of relatedness in
392 two datasets we examined are inversions. This may be common, but the example of
393 *Medicago truncatula* shows that polymorphic inversions are not ubiquitous. PCA has been
394 proposed as a method for discovering inversions (Ma and Amos 2012); however, the signal
395 left by inversions likely cannot be distinguished from long haplotypes under balancing
396 selection or simply regions of reduced recombination without additional lines of evidence.
397 Inversions show up in our method because across the inverted region, most gene trees
398 share a common split that dates back to the origin of the inversion. However, in many
399 applications, inversions are a nuisance. For instance, SMARTPCA (Patterson et al. 2006)
400 reduces their effect on PCA plots by regressing out the effect of linked SNPs on each
401 other. Removing samples with the less common orientation of each inversion reduced, but
402 did not eliminate, the signal of inversions seen in the *Drosophila melanogaster* dataset,
403 demonstrating that the genomic effects of transiently polymorphic inversions may outlast
404 the inversions themselves.

405 **The effect of selection** It seems that the variation in patterns of relatedness we see in
406 the *Medicago truncatula* and *Drosophila melanogaster* datasets must be explained some-
407 how by linked selection. Furthermore, the selection must be affecting many targets across
408 the genome, since we see similar effects across long distances (even distinct chromosomes).
409 For this reason, the most likely candidate may be selection against linked deleterious mu-
410 tations, known as “background selection” (Charlesworth et al. 1993; Charlesworth 2013).
411 Informally, background selection reduces the number of potential contributors to the gene
412 pool in regions of the genome with many possible deleterious mutations (Hudson and Ka-
413 plan 1995); for this reason, if it acts in a spatial context, it is expected to induce samples
414 from nearby locations to cluster together more frequently. Therefore, regions of the genome
415 harboring many targets of local adaptation may show similar patterns, since migrant alleles
416 in these regions will be selected against, and so locally gene trees will more closely reflect
417 spatial proximity.

418 A related possibility is that variation in patterns of relatedness is due to recent ad-
419 mixture between previously separated populations, the effects of which were not uniform
420 across the genome due to selection. For instance, it has been hypothesized that large-scale
421 variation in amount of introgressed Neanderthal DNA along the genome is due to selection
422 against Neanderthal genes, leading to greater introgression in regions of lower gene density
423 (Harris and Nielsen 2016; Juric et al. 2016). African *Drosophila melanogaster* are known to
424 have a substantial amount of recently introgressed genome from “cosmopolitan” sources;
425 if selection regularly favors genes from one origin, this could lead to substantial variation
426 in patterns of relatedness correlated with local gene density.

427 There has been substantial debate over the relative impacts of different forms of se-
428 lection. These have been difficult to disentangle in part because for the most part theory
429 makes predictions which are only strictly valid in randomly mating (i.e., unstructured)
430 populations, and it is unclear to what extent the spatial structure observed in most real

431 populations will affect these predictions. It may be possible to design more powerful statis-
432 tics that make stronger use of spatial information.

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

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

452 Finally, our software allows different choices for how many PCs to use in approximating
453 structure of each window (k in equation 1), and how many MDS coordinates to use when
454 describing the distance matrix between windows, but in our exploration, changing these has
455 not produced dramatically different results. These are all part of more general techniques
456 in dimension reduction and high-dimensional data visualization; we encourage the user to
457 experiment.

458 **Applications** So-called cryptic relatedness between samples has been one of the major
459 sources of confounding in genome-wide association studies (GWAS) and so methods must
460 account for it by modeling population structure or kinship (Astle and Balding 2009; Yang
461 et al. 2014). Since the effects of population structure is not constant along the genome, this
462 could in principle lead to an inflation of false positives in parts of the genome with stronger
463 population structure than the genome-wide average. A method such as ours might be used
464 to provide a more sensitive correction. Fortunately, in our human dataset this does not
465 seem likely to have a strong effect: most variation is due to small, independent regions,
466 possibly primarily inversions, and so may not have a major effect on GWAS. In the other
467 species we examined, particularly *Drosophila melanogaster*, treating population structure
468 as a single quantity would entail a substantial loss of power, and could potentially be

469 misleading.

470 **Acknowledgements**

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

477 **Disclosure declaration**

478 The authors declare no conflicts of interest.

479 **References**

- 480 Francesca Antonacci, Jeffrey M Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia
481 Siswara, Zhaoshi Jiang, and Evan E Eichler. Characterization of six human disease-
482 associated inversion polymorphisms. *Human molecular genetics*, 18(14):2555–2566, 2009.
- 483 William Astle and David J. Balding. Population structure and cryptic relatedness in genetic
484 association studies. *Statistical Science*, 24(4):451–471, 11 2009. doi: 10.1214/09-STS307.
485 URL <http://dx.doi.org/10.1214/09-STS307>.
- 486 Nicholas H. Barton. Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, 355(1403):
487 1553–1562, November 2000. doi: 10.1098/rstb.2000.0716. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1692896/>.
- 488 Albert P. Blair. Population structure in toads. *The American Naturalist*, 77(773):563–568,
489 1943. ISSN 00030147, 15375323. URL <http://www.jstor.org/stable/2457848>.
- 490 Yaniv Brandvain, Amanda M. Kenney, Lex Flagel, Graham Coop, and Andrea L. Sweigart.
491 Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genet*,
492 10(6):e1004410, 06 2014. doi: 10.1371/journal.pgen.1004410. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1004410>.
- 493 A. Brisbin, K. Bryc, J. Byrnes, F. Zakharia, L. Omberg, J. Degenhardt, A. Reynolds,
494 H. Ostrer, J. G. Mezey, and C. D. Bustamante. PCAdmix: principal components-based
495 assignment of ancestry along each chromosome in individuals with admixed ancestry
496 from two or more populations. *Hum. Biol.*, 84(4):343–364, August 2012.

- 499 K Bryc, A Auton, M R Nelson, J R Oksenberg, S L Hauser, S Williams, A Froment,
500 J M Bodo, C Wambebe, S A Tishkoff, and C D Bustamante. Genome-wide patterns of
501 population structure and admixture in West Africans and African Americans. *Proc Natl
502 Acad Sci U S A*, 107(2):786–791, January 2010. doi: 10.1073/pnas.0909559107. URL
503 <http://www.ncbi.nlm.nih.gov/pubmed/20080753?dopt=Abstract>.
- 504 R Burri, A Nater, T Kawakami, C F Mugal, P I Olason, L Smeds, A Suh, L Dutoit,
505 S Bureš, L Z Garamszegi, S Hogner, J Moreno, A Qvarnström, M Ružić, S A Sæther,
506 G P Sætre, J Török, and H Ellegren. Linked selection and recombination rate variation
507 drive the evolution of the genomic landscape of differentiation across the speciation
508 continuum of Ficedula flycatchers. *Genome Res*, 25(11):1656–1665, November 2015. doi:
509 10.1101/gr.196485.115. URL <https://www.ncbi.nlm.nih.gov/pubmed/26355005>.
- 510 Frank M.T.A. Busing, Erik Meijer, and Rien Van Der Leeden. Delete-m jackknife for
511 unequal m. *Statistics and Computing*, 9(1):3–8, 1999. ISSN 0960-3174. doi: 10.1023/A:
512 1008800423698. URL <http://dx.doi.org/10.1023/A%3A1008800423698>.
- 513 B Charlesworth, M T Morgan, and D Charlesworth. The effect of deleterious mutations
514 on neutral molecular variation. *Genetics*, 134(4):1289–1303, August 1993. URL <http://www.genetics.org/content/134/4/1289>.
- 516 Brian Charlesworth. Background selection 20 years on: The Wilhelmine E. Key 2012
517 invitational lecture. *Journal of Heredity*, 104(2):161–171, 2013. doi: 10.1093/jhered/
518 ess136. URL <http://jhered.oxfordjournals.org/content/104/2/161.abstract>.
- 519 Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton. The effects of genetic
520 and geographic structure on neutral variation. *Annual Review of Ecology, Evolution,
521 and Systematics*, 34(1):99–125, 2003. doi: 10.1146/annurev.ecolsys.34.011802.132359.
522 URL <http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.34.011802.132359>.
- 524 J M Comeron, R Ratnappan, and S Bailin. The many landscapes of recombination in
525 *Drosophila melanogaster*. *PLoS Genet*, 8(10), 2012. doi: 10.1371/journal.pgen.1002905.
526 URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3469467/>.
- 527 Russell B Corbett-Detig and Daniel L Hartl. Population genomics of inversion polymor-
528 phisms in *Drosophila melanogaster*. *PLoS Genet*, 8(12):e1003056, 2012.
- 529 T E Cruickshank and M W Hahn. Reanalysis suggests that genomic islands of speciation
530 are due to reduced diversity, not reduced gene flow. *Mol Ecol*, 23(13):3133–3157, 07 2014.
531 doi: 10.1111/mec.12796. URL <https://www.ncbi.nlm.nih.gov/pubmed/24845075>.
- 532 Nicolas Duforet-Frebourg, Keurcien Luu, Guillaume Laval, Eric Bazin, and Michael G.B.
533 Blum. Detecting genomic signatures of natural selection with principal component

- 534 analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*,
535 2015. doi: 10.1093/molbev/msv334. URL <http://mbe.oxfordjournals.org/content/early/2016/01/12/molbev.msv334.abstract>.
- 537 B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial
538 and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319. URL <http://pubs.siam.org/doi/abs/10.1137/1.9781611970319>.
- 540 Hans Ellegren, Linnea Smeds, Reto Burri, Pall I. Olason, Niclas Backstrom, Takeshi
541 Kawakami, Axel Kunstner, Hannu Makinen, Krystyna Nadachowska-Brzyska, Anna
542 Qvarnstrom, Severin Uebbing, and Jochen B. W. Wolf. The genomic landscape of species
543 divergence in Ficedula flycatchers. *Nature*, 491(7426):756–760, November 2012. ISSN
544 00280836. doi: 10.1038/nature11584. URL <http://dx.doi.org/10.1038/nature11584>.
- 545 Barbara E. Engelhardt and Matthew Stephens. Analysis of population structure: a uni-
546 fying framework and novel methods based on sparse factor analysis. *PLoS Genet*, 6(9),
547 September 2010. doi: 10.1371/journal.pgen.1001117. URL <http://www.ncbi.nlm.nih.gov/pubmed/20862358>.
- 549 Anna-Sophie Fiston-Lavier, Nadia D. Singh, Mikhail Lipatov, and Dmitri A. Petrov.
550 *Drosophila melanogaster* recombination rate calculator. *Gene*, 463(1–2):18 – 20, 2010.
551 ISSN 0378-1119. doi: <http://dx.doi.org/10.1016/j.gene.2010.04.015>. URL <http://www.sciencedirect.com/science/article/pii/S0378111910001769>.
- 553 B M Fitzpatrick, J R Johnson, D K Kump, J J Smith, S R Voss, and H B Shaffer.
554 Rapid spread of invasive genes into a threatened native species. *Proc Natl Acad Sci
555 U S A*, 107(8):3606–3610, February 2010. doi: 10.1073/pnas.0911802107. URL <http://www.ncbi.nlm.nih.gov/pubmed/20133596>.
- 557 Rafael F. Guerrero, François Rousset, and Mark Kirkpatrick. Coalescent patterns for
558 chromosomal inversions in divergent populations. *Philosophical Transactions of the
559 Royal Society B: Biological Sciences*, 367(1587):430–438, 2011. ISSN 0962-8436. doi:
560 10.1098/rstb.2011.0246. URL <http://rstb.royalsocietypublishing.org/content/367/1587/430>.
- 562 Kelley Harris and Rasmus Nielsen. The genetic cost of Neanderthal introgression. *Genetics*,
563 203(2):881–891, June 2016. URL <http://www.genetics.org/content/203/2/881>.
- 564 R R Hudson and N L Kaplan. Deleterious background selection with recombination. *Ge-
565 netics*, 141(4):1605–1617, December 1995. URL <http://www.genetics.org/content/141/4/1605>.
- 567 Emilia Huerta-Sánchez, Michael DeGiorgio, Luca Pagani, Ayele Tarekegn, Rosemary
568 Ekong, Tiago Antao, Alexia Cardona, Hugh E. Montgomery, Gianpiero L. Cavalleri,

- 569 Peter A. Robbins, Michael E. Weale, Neil Bradman, Endashaw Bekele, Toomas Kivisild,
570 Chris Tyler-Smith, and Rasmus Nielsen. Genetic signatures reveal high-altitude adapta-
571 tion in a set of Ethiopian populations. *Molecular Biology and Evolution*, 30(8):1877–1888,
572 2013. doi: 10.1093/molbev/mst089. URL <http://mbe.oxfordjournals.org/content/30/8/1877.abstract>.
- 574 Matthew B. Hufford, Pesach Lubinsky, Tanja Pyhäjärvi, Michael T. Devengenzo, Nor-
575 man C. Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgres-
576 sion in maize. *PLoS Genet*, 9(5):e1003477, 05 2013. doi: 10.1371/journal.pgen.1003477.
577 URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1003477>.
- 578 Ivan Juric, Simon Aeschbacher, and Graham Coop. The strength of selection against
579 Neanderthal introgression. *bioRxiv*, 2016. doi: 10.1101/030148. URL <http://biorxiv.org/content/early/2016/07/22/030148>.
- 581 N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis.
582 *Neural Computation*, 9(7):1493–1516, July 1997. ISSN 0899-7667. doi: 10.1162/neco.
583 1997.9.7.1493. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6795533.
- 585 Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking
586 along a recombining chromosome. *Genetics*, 160(2):765–777, 2002. URL <http://www.genetics.org/cgi/content/abstract/160/2/765>.
- 588 Mark Kirkpatrick. How and why chromosome inversions evolve. *PLoS Biol*, 8(9), 2010.
589 doi: 10.1371/journal.pbio.1000501. URL <http://www.ncbi.nlm.nih.gov/pubmed/20927412>.
- 591 Mark Kirkpatrick and Brian Barrett. Chromosome inversions, adaptive cassettes and the
592 evolution of species' ranges. *Molecular Ecology*, 2015. ISSN 1365-294X. doi: 10.1111/
593 mec.13074. URL <http://dx.doi.org/10.1111/mec.13074>.
- 594 Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-
595 Detig, Kristian A Stevens, Charles H Langley, and John E Pool. The Drosophila genome
596 nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including
597 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241, 2015.
- 598 C H Langley, K Stevens, C Cardeno, Y C Lee, D R Schrider, J E Pool, S A Langley,
599 C Suarez, R B Corbett-Detig, B Kolaczkowski, S Fang, P M Nista, A K Holloway, A D
600 Kern, C N Dewey, Y S Song, M W Hahn, and D J Begun. Genomic variation in natural
601 populations of *Drosophila melanogaster*. *Genetics*, 192(2):533–598, October 2012. doi:
602 10.1534/genetics.112.142018. URL <http://www.ncbi.nlm.nih.gov/pubmed/22673804>.

- 603 Thomas Lenormand. Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4):183 – 189, 2002. ISSN 0169-5347. doi: DOI:10.1016/S0169-5347(02)02497-7. URL <http://www.sciencedirect.com/science/article/pii/S0169534702024977>.
- 607 J Ma and C I Amos. Investigation of inversion polymorphisms in the human genome
608 using principal components analysis. *PLoS One*, 7(7), 2012. doi: 10.1371/journal.pone.
609 0040224. URL <http://www.ncbi.nlm.nih.gov/pubmed/22808122>.
- 610 Trudy F. C. Mackay, Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F.
611 Ayroles, Dianhui Zhu, Sonia Casillas, Yi Han, Michael M. Magwire, Julie M. Crid-
612 land, Mark F. Richardson, Robert R. H. Anholt, Maite Barron, Crystal Bess, Ker-
613 stin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura
614 Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N. Jhangiani,
615 Katherine W. Jordan, Fremiet Lara, Faye Lawrence, Sandra L. Lee, Pablo Librado,
616 Raquel S. Linheiro, Richard F. Lyman, Aaron J. Mackey, Mala Munidasa, Donna Marie
617 Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel
618 Ramia, Jeffrey G. Reid, Stephanie M. Rollmann, Julio Rozas, Nehad Saada, Lavanya
619 Turlapati, Kim C. Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M.
620 Bergman, Kevin R. Thornton, David Mittelman, and Richard A. Gibbs. The *Drosophila*
621 *melanogaster* genetic reference panel. *Nature*, 482(7384):173–178, February 2012. ISSN
622 00280836. URL <http://dx.doi.org/10.1038/nature10811>.
- 623 José V Manjón, Pierrick Coupé, Luis Concha, Antonio Buades, D Louis Collins, and
624 Montserrat Robles. Diffusion weighted image denoising using overcomplete local PCA.
625 *PloS one*, 8(9):e73021, 2013.
- 626 Simon Henry Martin, Markus Moest, William J Palmer, Camilo Salazar, W. Owen McMil-
627 lan, Francis M Jiggins, and Chris D Jiggins. Natural selection and genetic diversity
628 in the butterfly *Heliconius melpomene*. *Genetics*, 203(1):525–541, May 2016. doi:
629 10.1101/042796. URL <http://www.genetics.org/content/203/1/525>.
- 630 Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*,
631 5(10):e1000686, 2009.
- 632 P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in
633 Europeans. *Science*, 201(4358):786–792, September 1978. URL <http://www.ncbi.nlm.nih.gov/pubmed/356262>.
- 635 N J Nadeau, A Whibley, R T Jones, J W Davey, K K Dasmahapatra, S W Baxter,
636 M A Quail, M Joron, R H ffrench Constant, M L Blaxter, J Mallet, and C D Jiggins.
637 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-
638 scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci*, 367(1587):343–353,

- 639 February 2012. doi: 10.1098/rstb.2011.0198. URL <http://www.ncbi.nlm.nih.gov/pubmed/22201164>.
- 640
- 641 M R Nelson, K Bryc, K S King, A Indap, A R Boyko, J Novembre, L P Briley, Y Maruyama,
642 D M Waterworth, G Waeber, P Vollenweider, J R Oksenberg, S L Hauser, H A Stirnadel,
643 J S Kooner, J C Chambers, B Jones, V Mooser, C D Bustamante, A D Roses, D K Burns,
644 M G Ehm, and E H Lai. The Population Reference Sample, POPRES: a resource for
645 population, disease, and pharmacological genetics research. *Am J Hum Genet*, 83(3):
646 347–358, September 2008. doi: 10.1016/j.ajhg.2008.08.005. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556436/?tool=pubmed>.
- 647
- 648 John Novembre and Matthew Stephens. Interpreting principal component analyses of
649 spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.
- 650 John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam
651 Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes
652 mirror geography within europe. *Nature*, 456(7218):98–101, 2008.
- 653 Timothy Paape, Peng Zhou, Antoine Branca, Roman Briskine, Nevin Young, and Peter
654 Tiffin. Fine-scale population recombination rates, hotspots, and correlates of recombina-
655 tion in the *Medicago truncatula* genome. *Genome Biology and Evolution*, 4(5):726–737,
656 2012. doi: 10.1093/gbe/evs046. URL <http://gbe.oxfordjournals.org/content/4/5/726.abstract>.
- 657
- 658 Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis.
659 *PLoS Genetics*, 2(12):e190, 12 2006. doi: 10.1371/journal.pgen.0020190. URL <http://dx.plos.org/10.1371%2Fjournal.pgen.0020190>.
- 660
- 661 J B Pease and M W Hahn. More accurate phylogenies inferred from low-recombination
662 regions in the presence of incomplete lineage sorting. *Evolution*, 67(8):2376–2384, August
663 2013. doi: 10.1111/evo.12118. URL <http://www.ncbi.nlm.nih.gov/pubmed/23888858>.
- 664 John E Pool. The mosaic ancestry of the Drosophila Genetic Reference Panel and the
665 *D. melanogaster* reference genome reveals a network of epistatic fitness interactions.
666 *Molecular Biology and Evolution*, 32(12):3236–3251, 2015. doi: 10.1101/014837. URL
667 <http://mbe.oxfordjournals.org/content/32/12/3236.abstract>.
- 668
- 669 John E. Pool, Russell B. Corbett-Detig, Ryuichi P. Sugino, Kristian A. Stevens, Charis M.
670 Cardeno, Marc W. Crepeau, Pablo Duchen, J. J. Emerson, Perot Saelao, David J. Begun,
671 and Charles H. Langley. Population genomics of sub-Saharan *Drosophila melanogaster*:
672 African diversity and non-African admixture. *PLoS Genet*, 8(12):1–24, 12 2012. doi:
673 10.1371/journal.pgen.1003080. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1003080>.

- 674 Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick,
675 and David Reich. Principal components analysis corrects for stratification in genome-
676 wide association studies. *Nature genetics*, 38(8):904–909, 2006.
- 677 Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems*,
678 2016. URL <https://CRAN.R-project.org/package=RSpectra>. R package version 0.11-
679 0.
- 680 Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear
681 embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075. doi: 10.1126/science.
682 290.5500.2323. URL <http://science.sciencemag.org/content/290/5500/2323>.
- 683 Fabian Staubach, Anna Lorenc, Philipp W. Messer, Kun Tang, Dmitri A. Petrov, and
684 Diethard Tautz. Genome patterns of selection and introgression of haplotypes in natural
685 populations of the house mouse (*Mus musculus*). *PLoS Genet*, 8(8):e1002891, 08 2012.
686 doi: 10.1371/journal.pgen.1002891. URL <http://dx.doi.org/10.1371%2Fjournal.pgen.1002891>.
- 688 Haibao Tang, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo
689 Zhou, Laurent Gentzbittel, Kevin L Childs, Mark Yandell, Heidrun Gundlach, et al. An
690 improved genome release (version mt4. 0) for the model legume *Medicago truncatula*.
691 *BMC genomics*, 15(1):1, 2014.
- 692 T L Turner, M W Hahn, and S V Nuzhdin. Genomic islands of speciation in *Anopheles
693 gambiae*. *PLoS Biol*, 3(9), September 2005. doi: 10.1371/journal.pbio.0030285. URL
694 <https://www.ncbi.nlm.nih.gov/pubmed/16076241>.
- 695 Benjamin Vernot and Joshua M. Akey. Resurrecting surviving neandertal lineages from
696 modern human genomes. *Science*, 2014. doi: 10.1126/science.1245938. URL <http://www.sciencemag.org/content/early/2014/01/28/science.1245938.abstract>.
- 698 Ian J. Wang and Gideon S. Bradburd. Isolation by environment. *Molecular Ecology*, 23
699 (23):5649–5662, 2014. ISSN 1365-294X. doi: 10.1111/mec.12938. URL <http://dx.doi.org/10.1111/mec.12938>.
- 701 Andreas Weingessel and Kurt Hornik. Local PCA algorithms. *Neural Networks, IEEE
702 Transactions on*, 11(6):1242–1250, 2000.
- 703 Sewall Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354,
704 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>.
- 706 Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price.
707 Advantages and pitfalls in the application of mixed-model association methods. *Nat*

710 A Choosing window length

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

718 Since we summarize patterns of relatedness using relative positions in the principal
 719 component maps, we quantify “noise” as the standard error of a sample’s position on PC1
 720 in a particular window, averaged across windows and samples, and “signal” as the standard
 721 deviation of the sample’s position on PC1 over all windows, averaged over samples. The
 722 definition of eigenvectors does not specify their sign, and so when comparing between
 723 windows we choose signs to best match each other: after choosing $PC1_1$, for instance,
 724 if u is the first eigenvector obtained from the covariance matrix for window j , then we
 725 next choose $PC1_j = \pm u$, where the sign is chosen according to which of $\|PC1_1 - u\|$ or
 726 $\|PC1_1 + u\|$ is smaller.

727 After doing this, the mean variance across windows is

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

728 where $PC1_{ij}$ is the position of the i^{th} individual on $PC1$ in window j , and $\overline{PC1}_j =$
 729 $(1/N) \sum_{j=1}^N PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jack-
 730 knife (Busing et al. 1999; Efron 1982): we divide the j^{th} window into 10 equal-sized
 731 pieces, and let $PC1_{ij,k}$ denote the first principal component of this region found af-
 732 ter removing the k^{th} piece; then the estimate of the squared standard error is $\sigma_{ij}^2 =$
 733 $\frac{9}{10} \sum_{k=1}^{10} (PC1_{ij,k} - \frac{1}{10} \sum_{\ell=1}^{10} PC1_{ij,\ell})^2$. Averaging over samples and windows,

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

734 For the main analysis, we defined windows to each consist of the same number of neigh-
 735 boring SNPs, and calculated σ_{signal}^2 and σ_{noise}^2 for a range of window sizes (i.e., numbers

736 of SNPs). For our main results we chose the smallest window for which σ_{signal}^2 was con-
 737 sistently larger than σ_{noise}^2 (but checked other sizes); the values for various window sizes
 738 across *Drosophila* chromosomes are shown in Table S1. In the cases we examined, we found
 739 nearly identical results after varying window size, and choosing windows to be of the same
 740 physical length (in bp) rather than in numbers of SNPs.

		window length (SNPs)				
chrom.	arm	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).

741 B Weighted PCA

742 Principal components analysis can be thought of as finding a good low-dimensional matrix
 743 factorization (Engelhardt and Stephens 2010) that well-approximates the original data in
 744 the least-squares sense: if C is the $N \times N$ genetic covariance matrix, then to find the top k
 745 principal components, we find an orthogonal $N \times k$ matrix U , and a $k \times k$ diagonal matrix
 746 Λ with diagonal entries $\Lambda_{ii} = \lambda_i$ to minimize

$$\|C - U\Lambda U^T\|^2 = \sum_{ij} \left(C_{ij} - \sum_m \lambda_m U_{im} U_{jm} \right)^2. \quad (3)$$

747 The columns of U , known as the principal components, are the eigenvectors of C , the
 748 entries of λ are the eigenvalues of C , and the proportion of variance explained by the m^{th}

749 component is

$$\frac{\lambda_m^2}{\sum_\ell \lambda_\ell^2} = \frac{\sum_{ij} (\lambda_m U_{im} U_{jm})^2}{\sum_{ij} C_{ij}^2}.$$

750 Thinking about the problem as a least-squares approximation problem makes it clear
 751 why unbalanced sample sizes can result in undesirable outcomes. If we want to describe
 752 variation *between* populations, but 80% of the samples are from a single population, then
 753 unless populations are highly differentiated, a better approximation to C may be obtained
 754 by using the columns of U to describe variation *within* the overrepresented population
 755 rather than between the populations. A common workaround is to remove samples, but a
 756 more elegant solution can be found by reweighting the objective function in (3). Let w_i be
 757 a weight associated with sample i , W the diagonal matrix with w along the diagonal, and
 758 instead seek to minimize

$$\|W^{1/2}(C - U\Lambda U^T)W^{1/2}\|^2 = \sum_{ij} w_i w_j \left(G_{ij} - \sum_m \lambda_m U_{im} U_{jm} \right)^2, \quad (4)$$

759 and now for convenience we require U to be orthogonal in $\ell_2(w)$, i.e., that $U^T W U = I$.
 760 We then would choose w to give roughly equal weight to each *population*, instead of each
 761 individual. We have used with good results the weightings $w_i = 1/\max(10, n_i)$, where
 762 n_i is, if there are discrete populations, the number of samples in the same population as
 763 sample i ; or, for continuously sampled individuals, the number of samples within a certain
 764 distance of sample i .

765 To solve (4), let λ and V denote the top k eigenvalues and eigenvectors of $W^{1/2}CW^{1/2}$,
 766 so that $V\Lambda V^T$ is the rank k matrix closest in least squares to $W^{1/2}CW^{1/2}$; so if we define
 767 $U = W^{-1/2}V$ then $U^T W U = V^T V = I$, and

$$W^{-1/2}V\Lambda V^T W^{-1/2} = U\Lambda U^T$$

768 is the low-dimensional approximation to C . The proportion of variance explained is calculated
 769 from eigenvalues as before, but has the interpretation

$$\frac{\lambda_m^2}{\sum_\ell \lambda_\ell^2} = \frac{\sum_{ij} w_i w_j (\lambda_m U_{im} U_{jm})^2}{\sum_{ij} w_i w_j C_{ij}^2}.$$

770 In our R implementation we use the Spectra library (Qiu and Mei 2016) to find only the
 771 top k eigenvectors.

772 **Testing** To demonstrate the utility of this method XXX see Figure S25.

	10000 SNPs MDS1	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: C

orrelations 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 [XXX](#)) 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. Plots of MDS values along the genome are visually nearly identical for parameter sets having similar window sizes – full reports are available as supplementary material on Data Dryad [XXX](#).

⁷⁷³ **C Supplementary Tables**

⁷⁷⁴ **D Supplementary Figures**

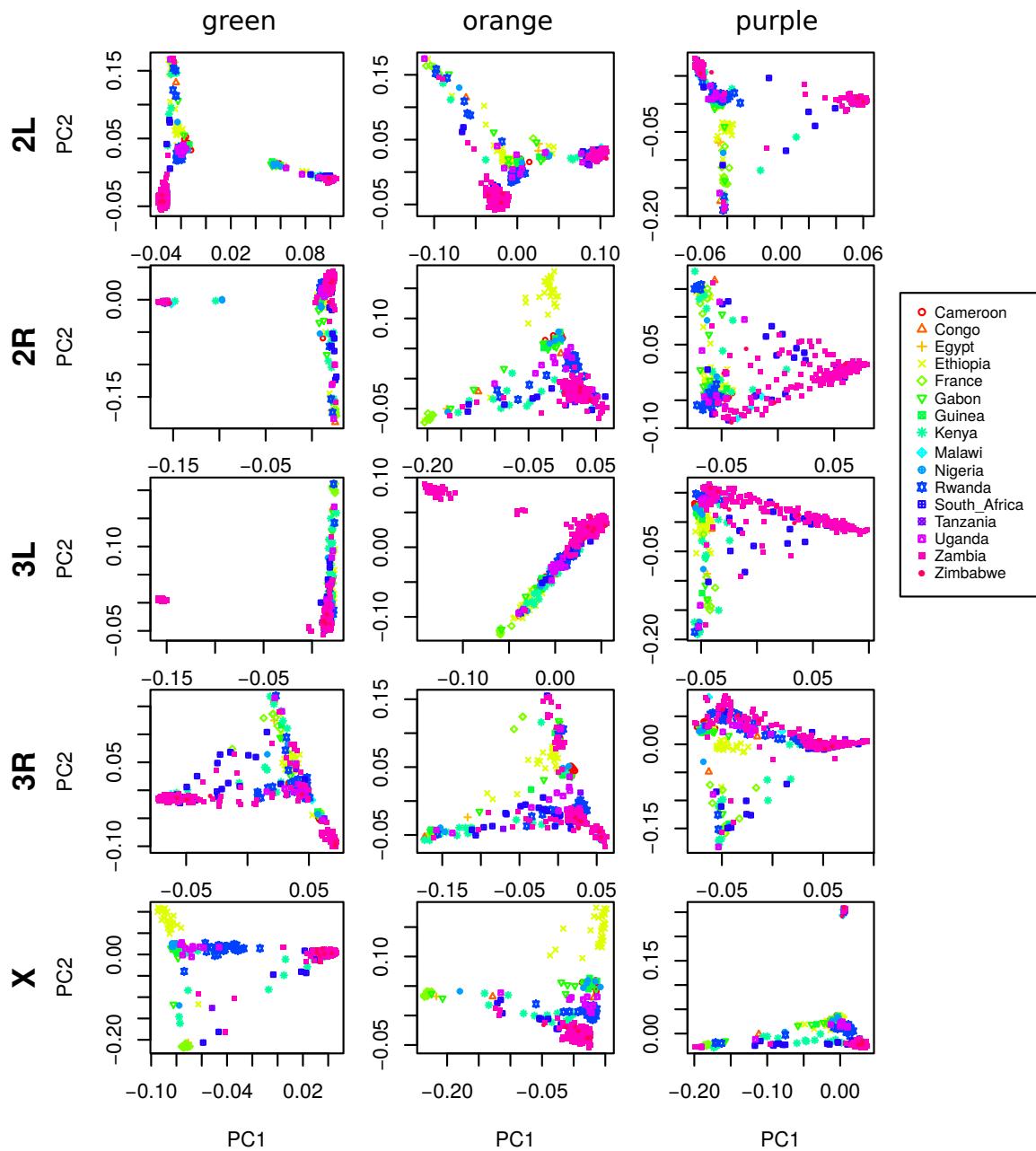


Figure S1: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange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

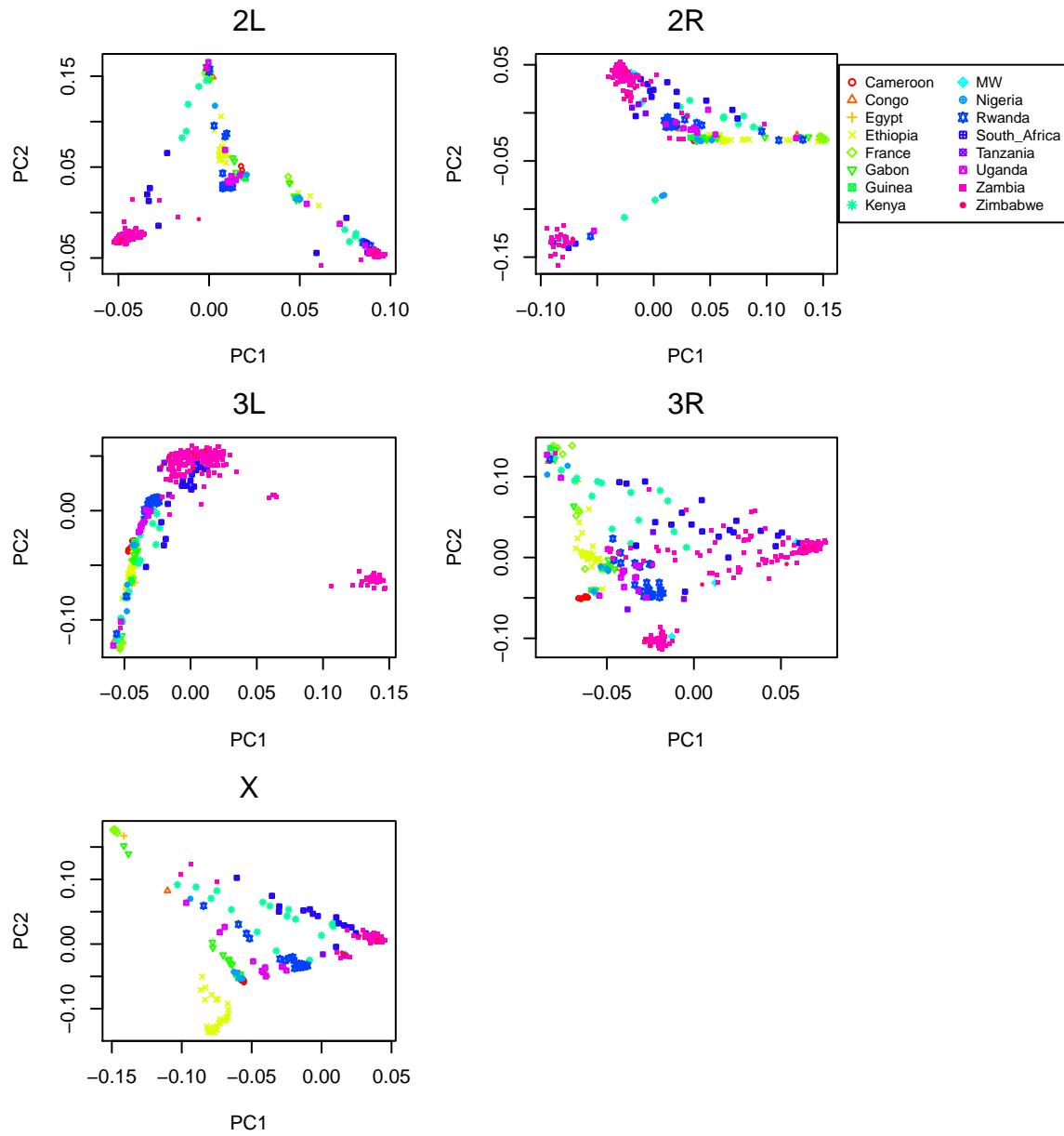


Figure S2: PCA plots for chromosome arms 2L, 2R, 3L, 3R and X of the *Drosophila melanogaster* dataset.

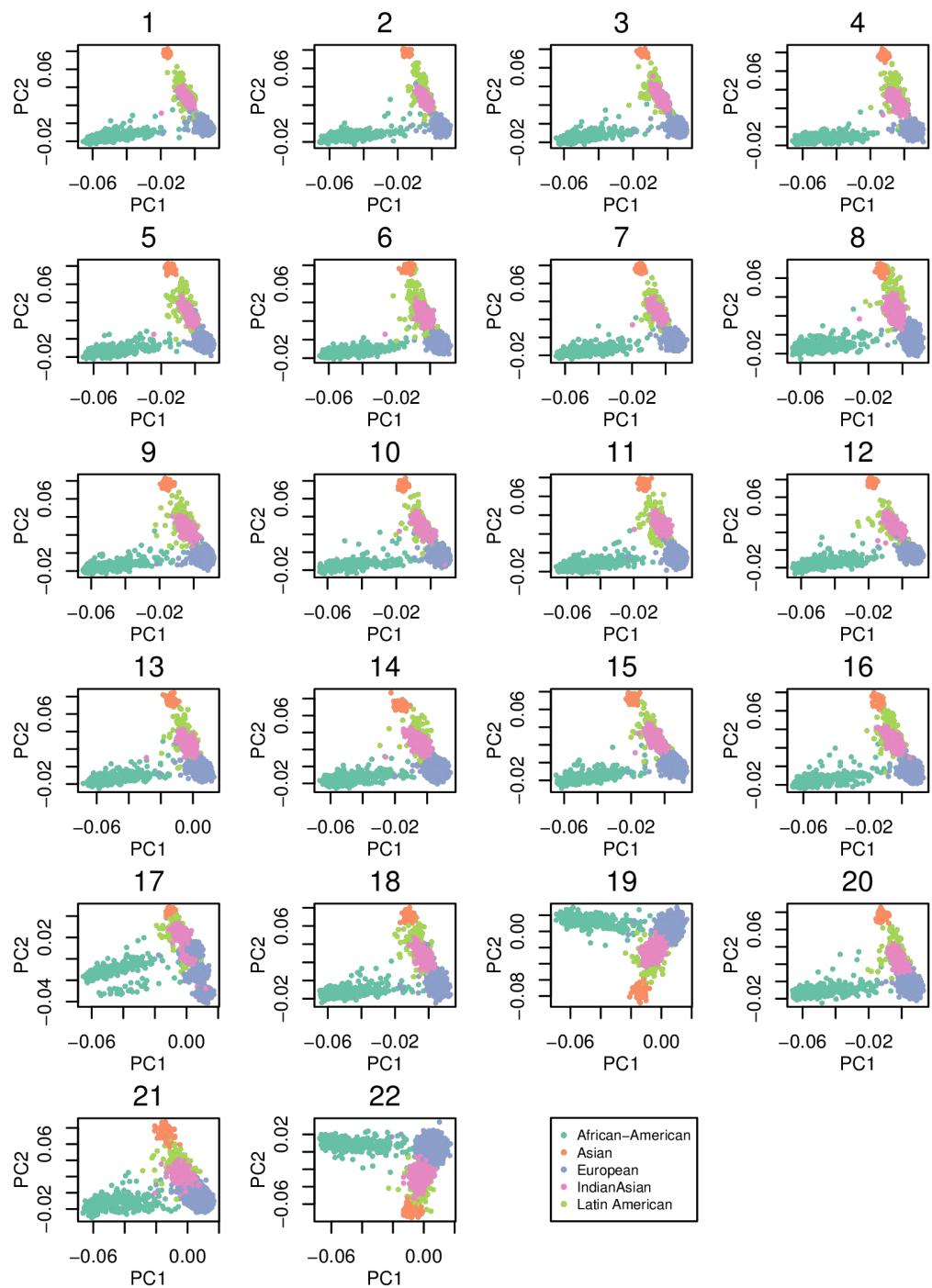


Figure S3: PCA plots for all 22 human autosomes from the POPRES data.

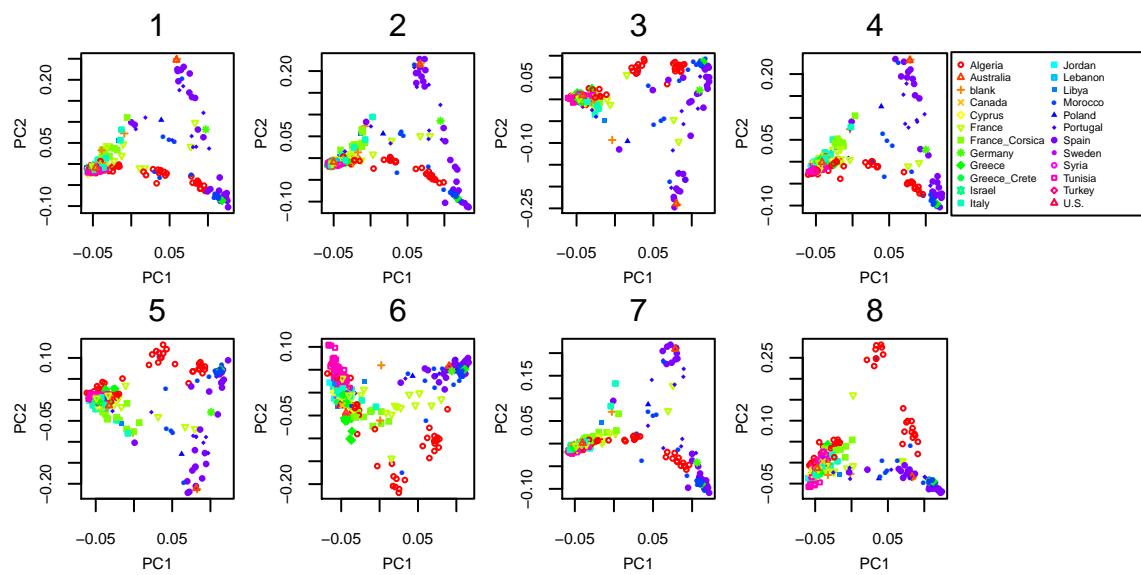


Figure S4: PCA plots for all 8 chromosomes in the *Medicago truncatula* dataset.

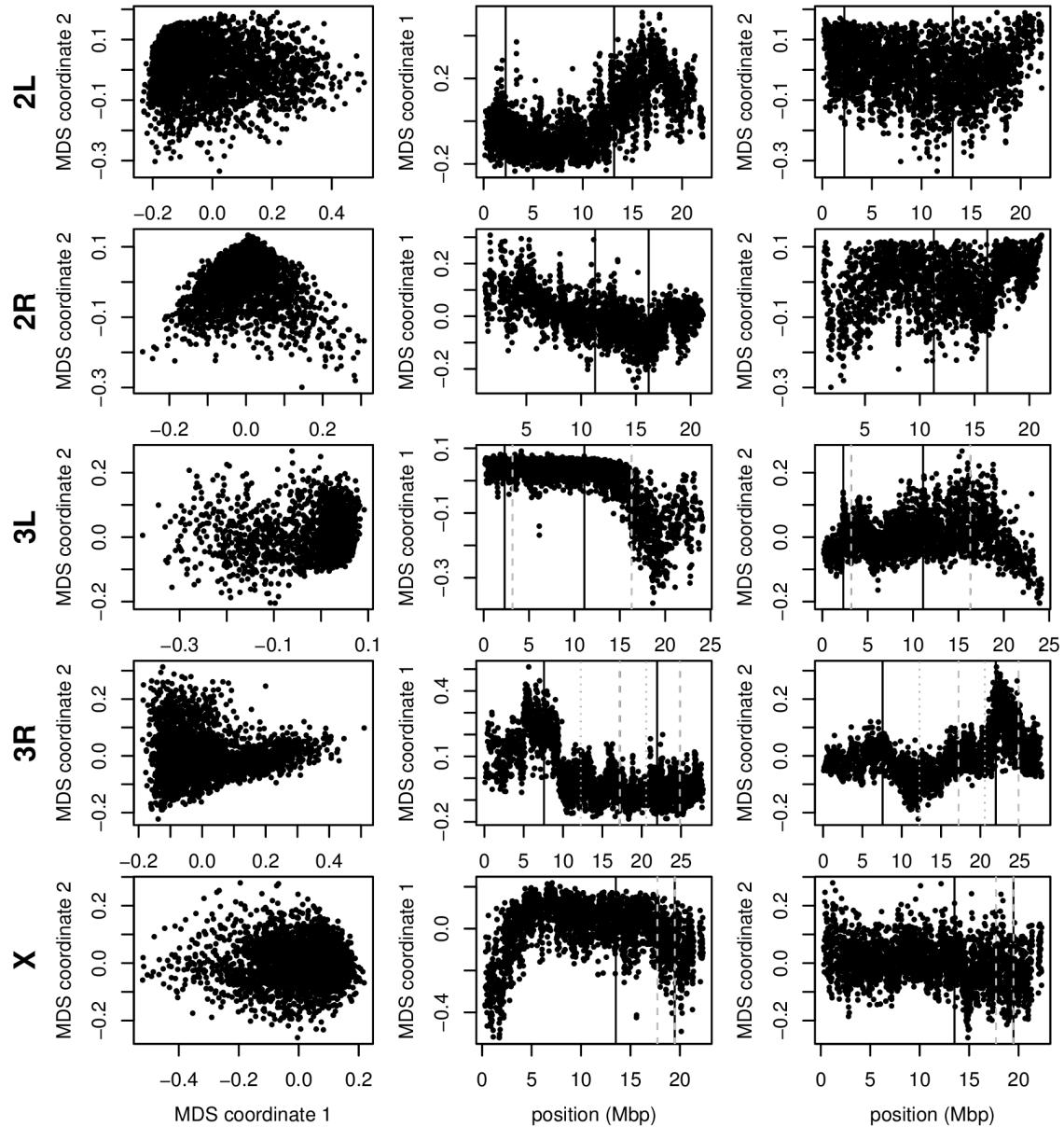


Figure S5: Variation in structure for windows of 1,000 SNPs across *Drosophila melanogaster* chromosome arms: without inversions. As in Figure 2, but after omitting for each chromosome arm individuals carrying the less frequent orientation of any inversions on that chromosome arm. The values differ from 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.

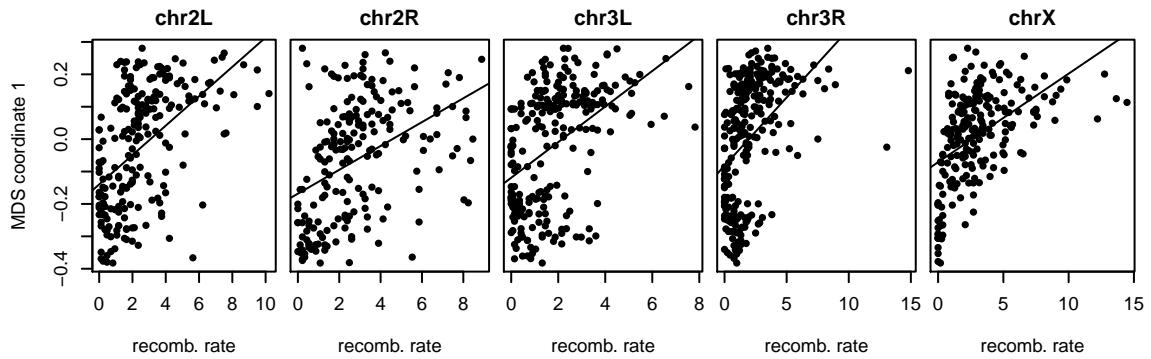


Figure S6: Recombination rate, and the effects of population structure for *Drosophila melanogaster*: this shows the first MDS coordinate and recombination rate (in cM/Mbp), as in Figure 4, against each other. Since the windows underlying estimates of Figure 4 do not coincide, to obtain correlations 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, $r^2 = 0.24$.

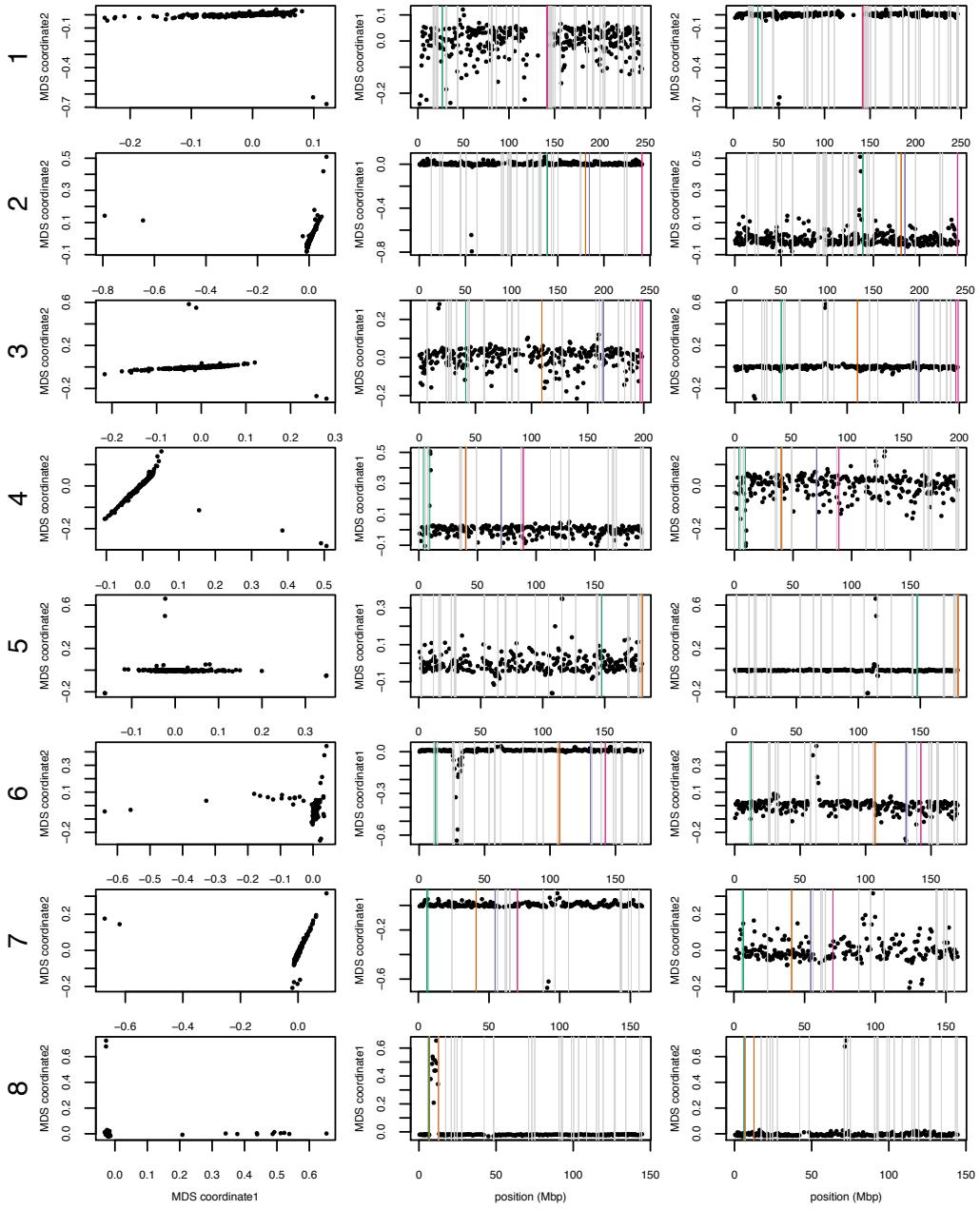


Figure S7: MDS plots for human chromosomes 1-8. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosomes. Colorful vertical lines show the breakpoints of known valid inversions, while grey vertical lines show the breakpoints of predicted inversions.

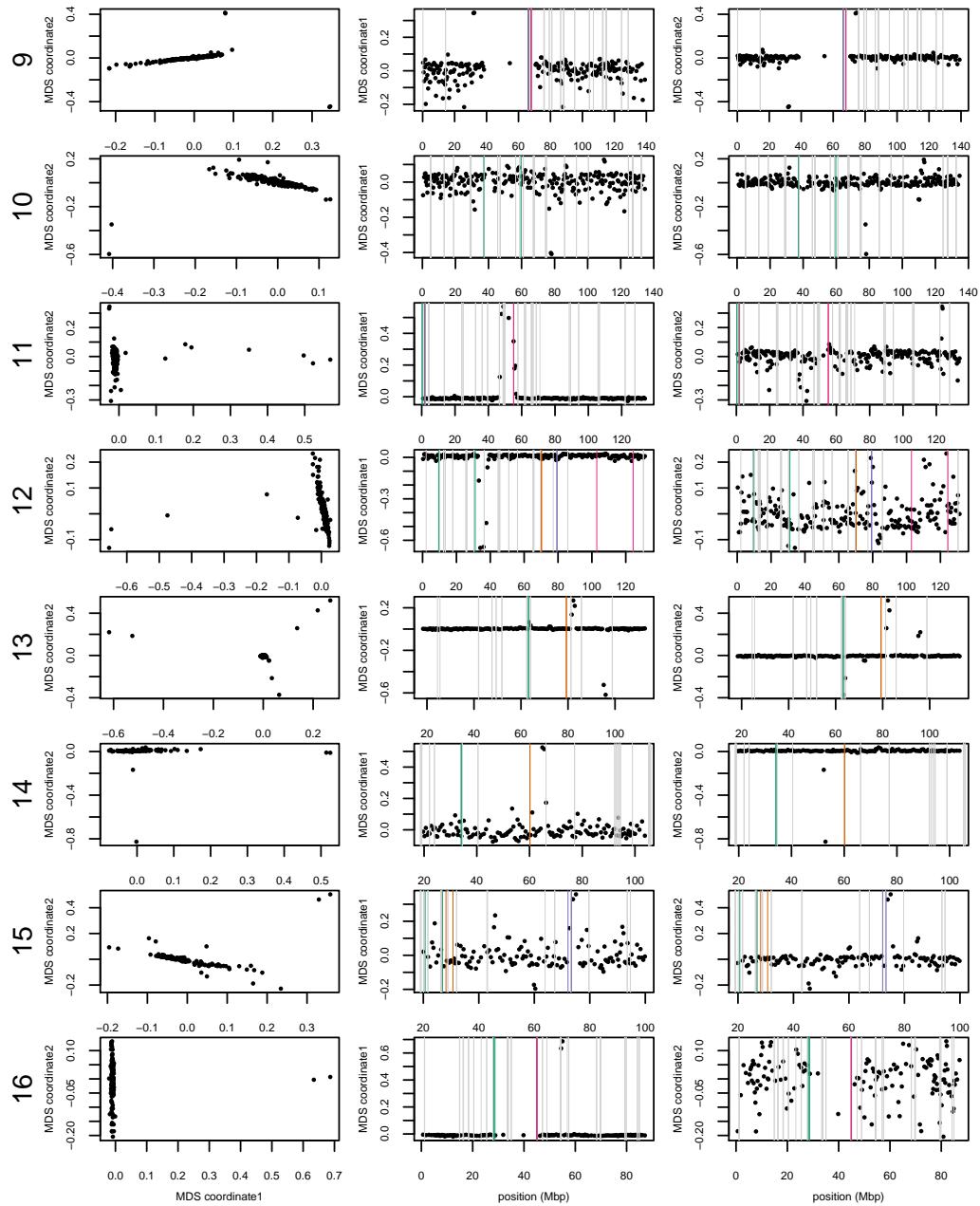


Figure S8: MDS plots for human chromosomes 9-16, as in Supplemental Figure S7.

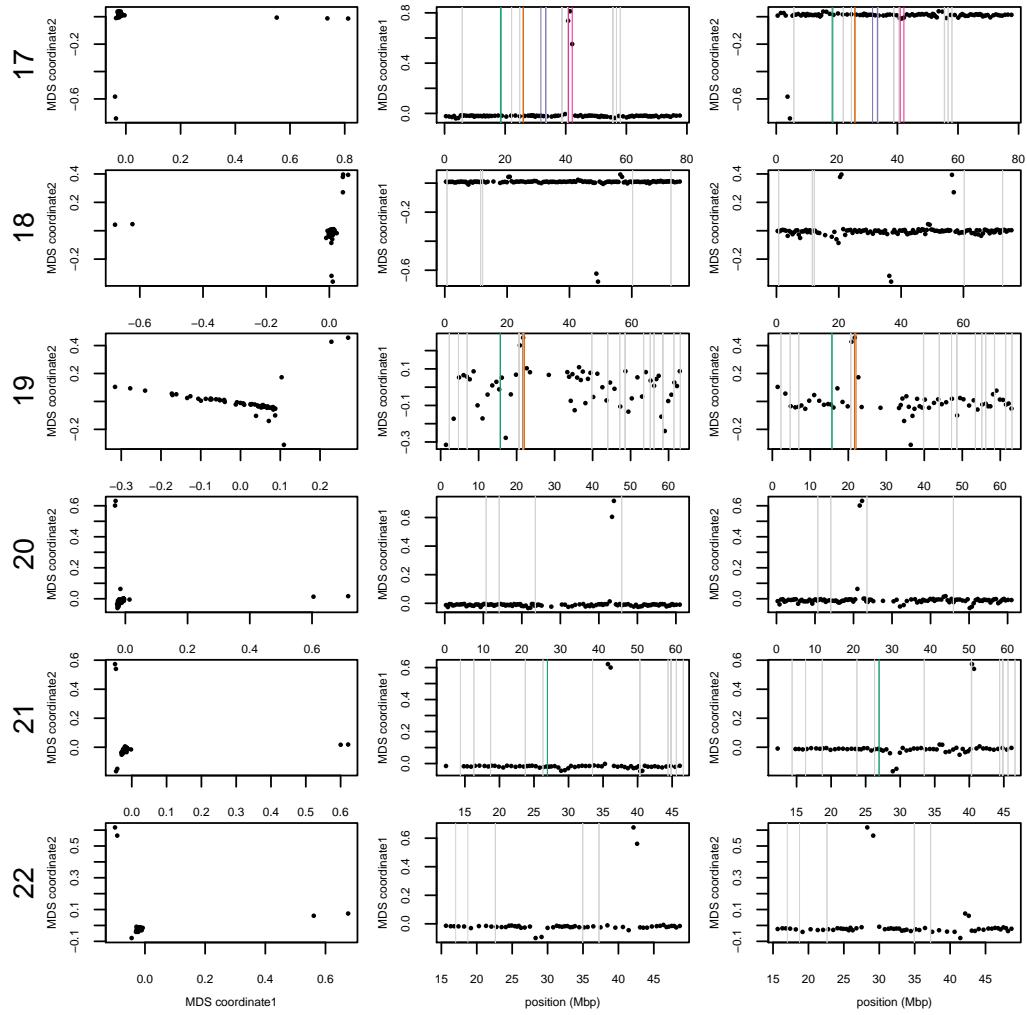


Figure S9: MDS plots for human chromosomes 17-22, as in Supplemental Figure S7.

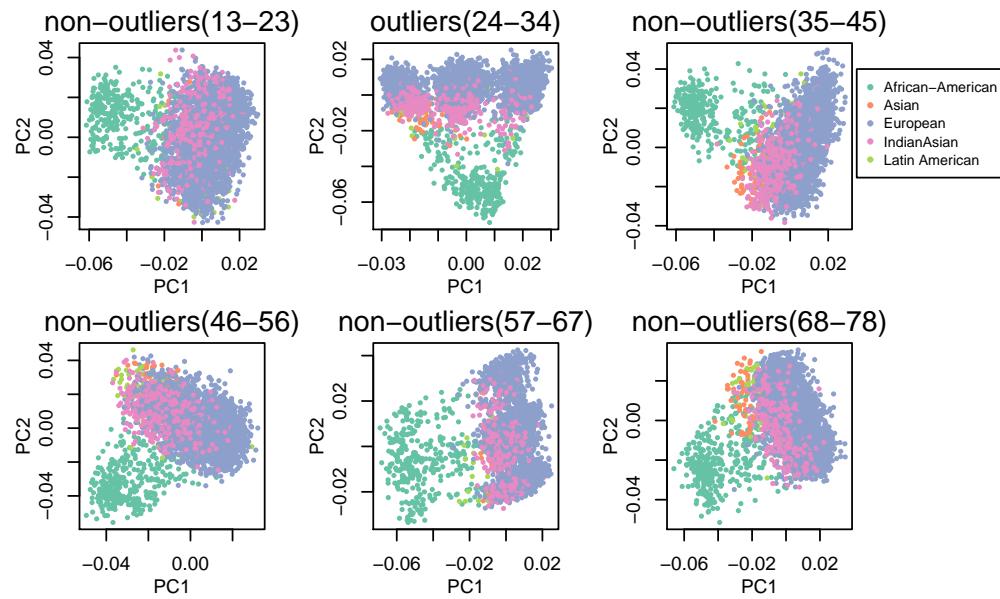


Figure S10: Comparison of PCA figures within outlying windows (center column) and flanking non-outlying windows (left and right columns) for the two windows having outlying MDS scores on chromosome 8.

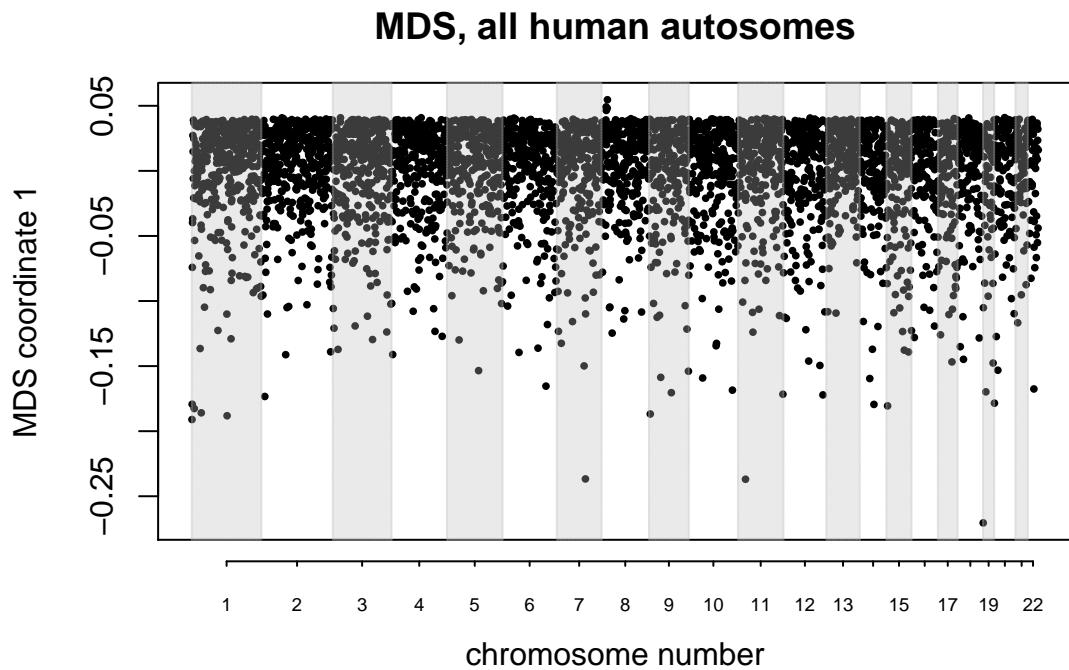


Figure S11: MDS visualization of variation in the effects of population structure amongst windows across *all* human autosomes simultaneously. The small group of windows with positive MDS values lie around the inversion at 8p23.

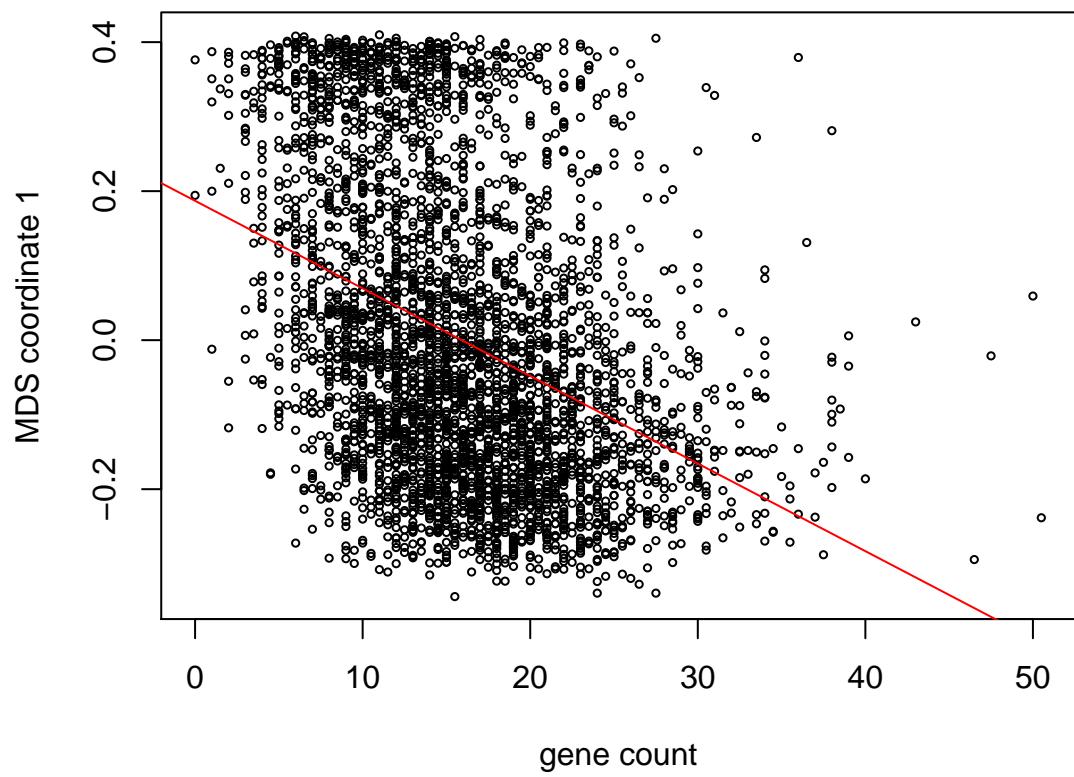


Figure S12: First MDS coordinate against gene density for all 8 chromosomes of *M. truncatula*. The first MDS coordinate is significantly correlated with gene count ($r = 0.149$, $p = 2.2 \times 10^{-16}$).

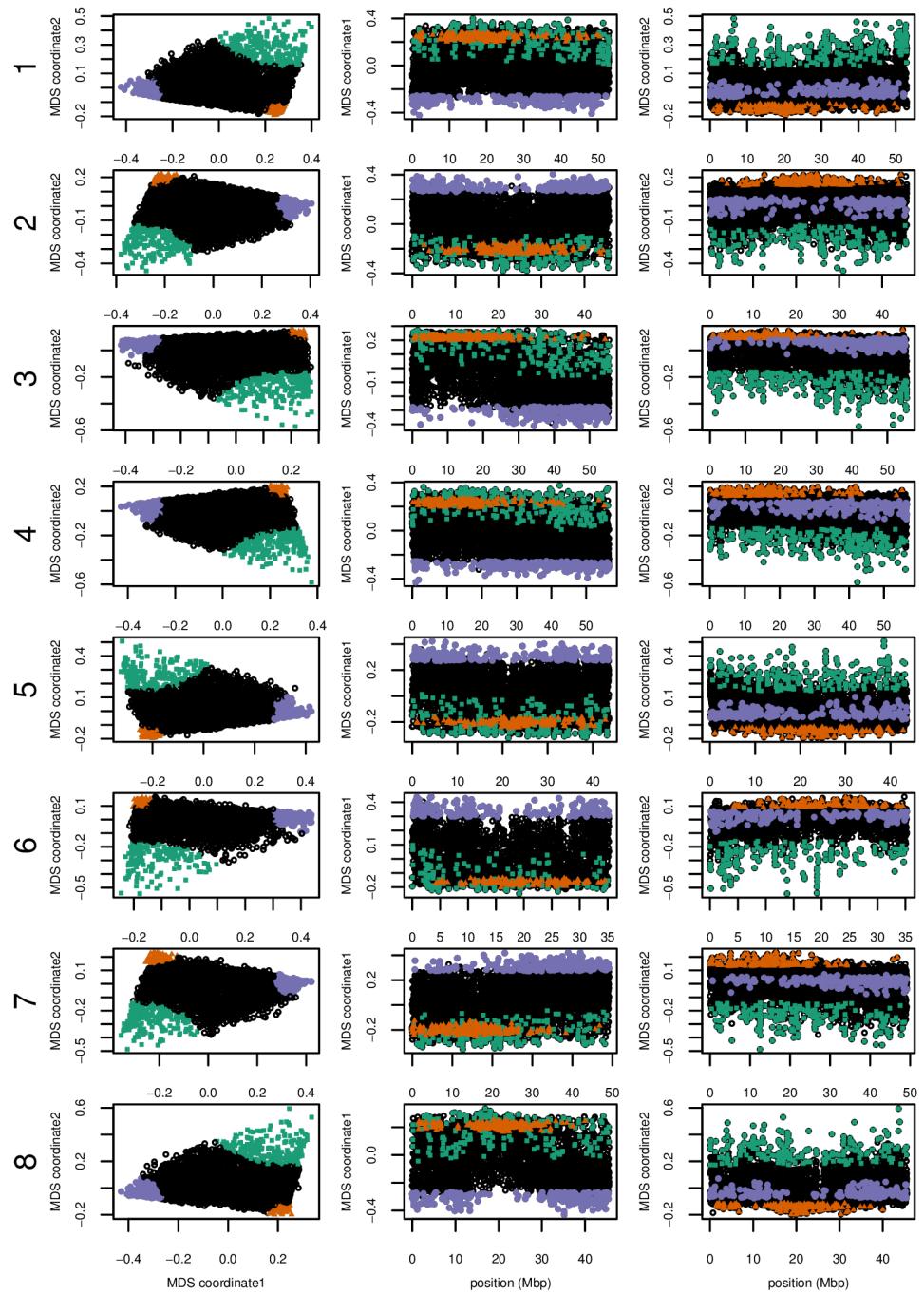


Figure S13: MDS visualizations of the effects of population structure for all 8 chromosomes of the *Medicago truncatula* data.

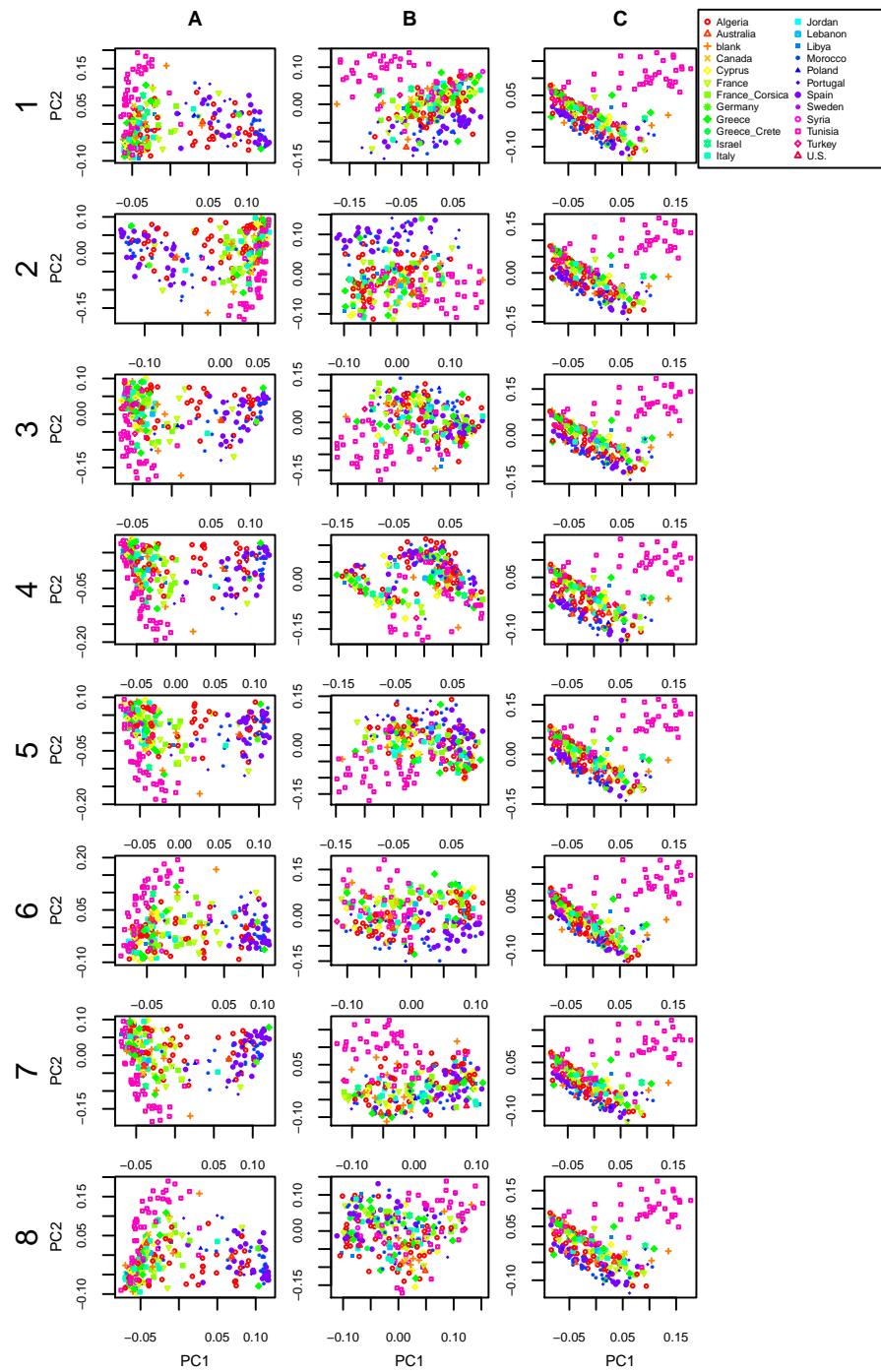


Figure S14: PCA plots for regions colored in Figure S13 on all 8 chromosomes of *Medicago truncatula*: (A) green, (B) orange, and (C) purple.

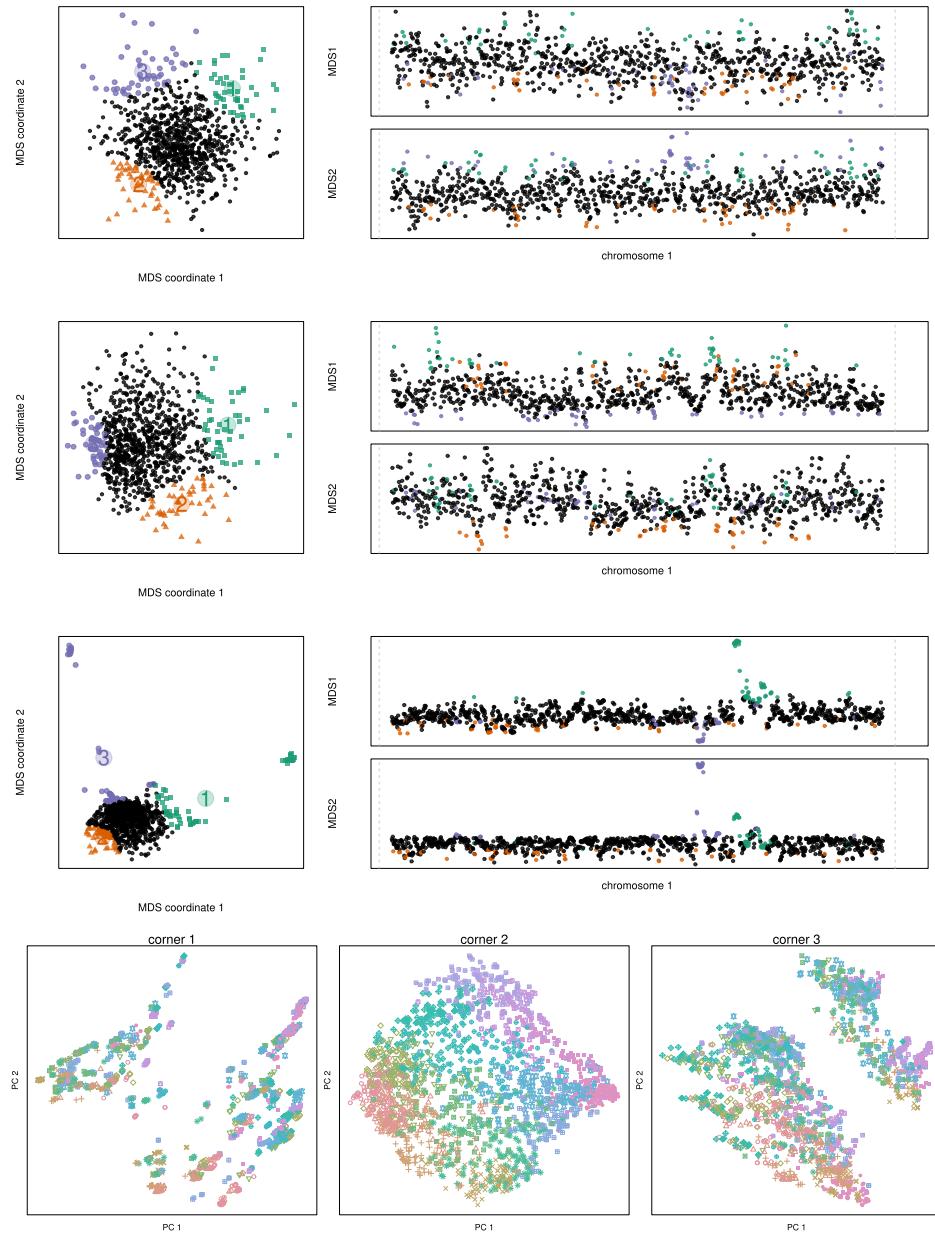


Figure S15: MDS visualizations of the results of individual-based simulations using SLiM (see Appendix ?? for details). All simulations are neutral, and recombination is: **(top)** constant; **(top middle)** varies stepwise by factors of two in seven equal-length segments, with highest rates on the ends, 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 ⁴⁶ 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 S16: PCA plots for regions colored in Figure ??

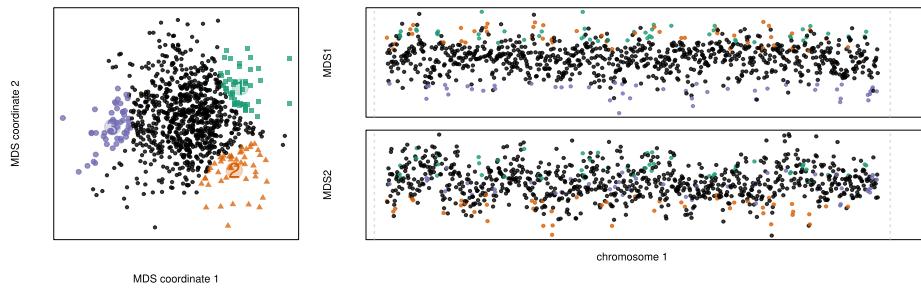


Figure S17: MDS visualizations of simulation run 005464 (RECOMBTYPE, “base”; SELTYPE, “selected”)

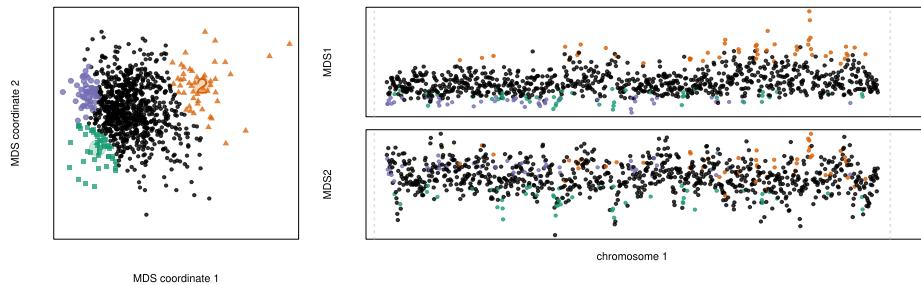


Figure S18: *KEEP* MDS visualizations of simulation run 031486 (RECOMBTYPE, “base”; SELTYPE, “selected”)

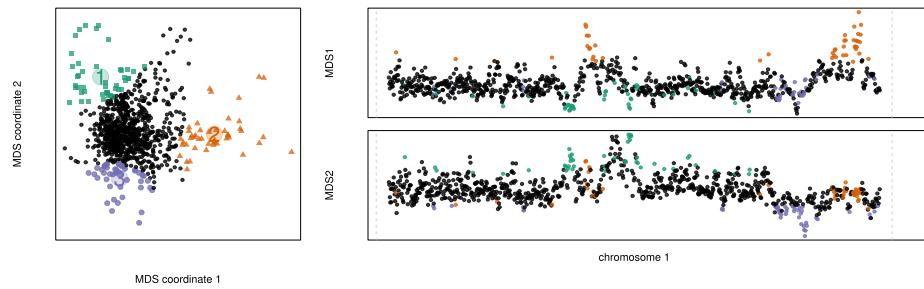


Figure S19: MDS visualizations of simulation run 009673 (RECOMBTYPE, “step”; SELTYPE, “selected”)

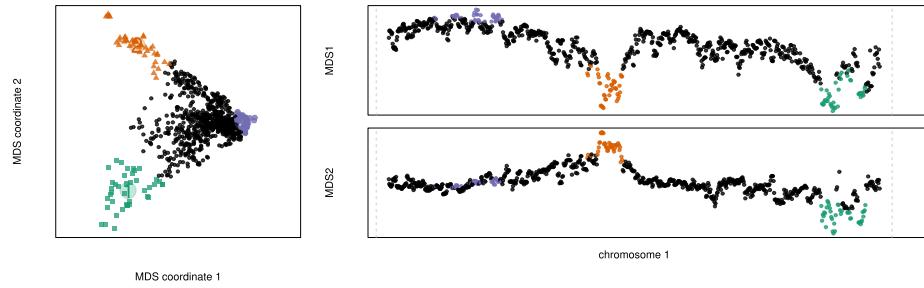


Figure S20: MDS visualizations of simulation run 009673 (RECOMBTYPE, “base”; SELTYPE, “balancing”)

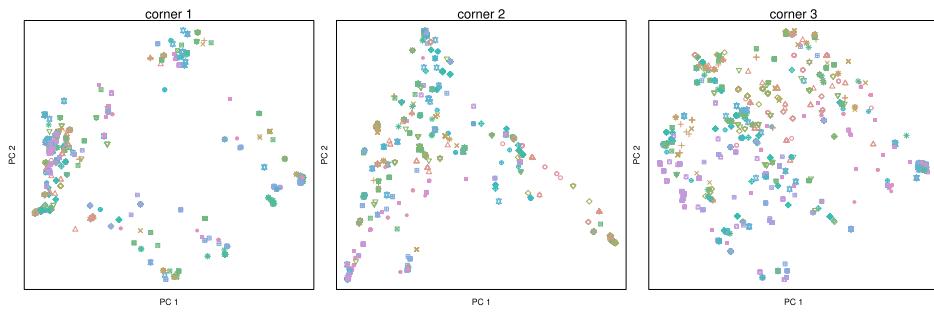


Figure S21: PCA plots for regions colored in Figure S20

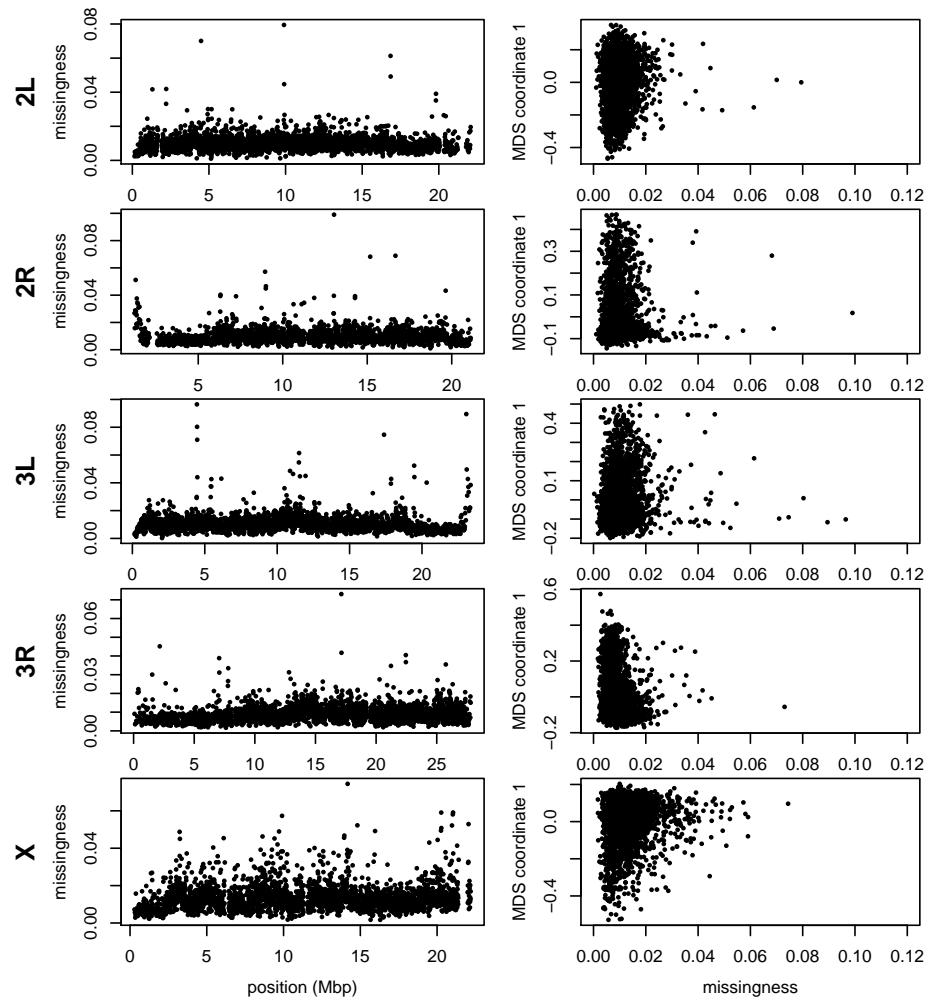


Figure S22: The missingness ratio along genome and the correlation of MDS against missingness in *Drosophila*.

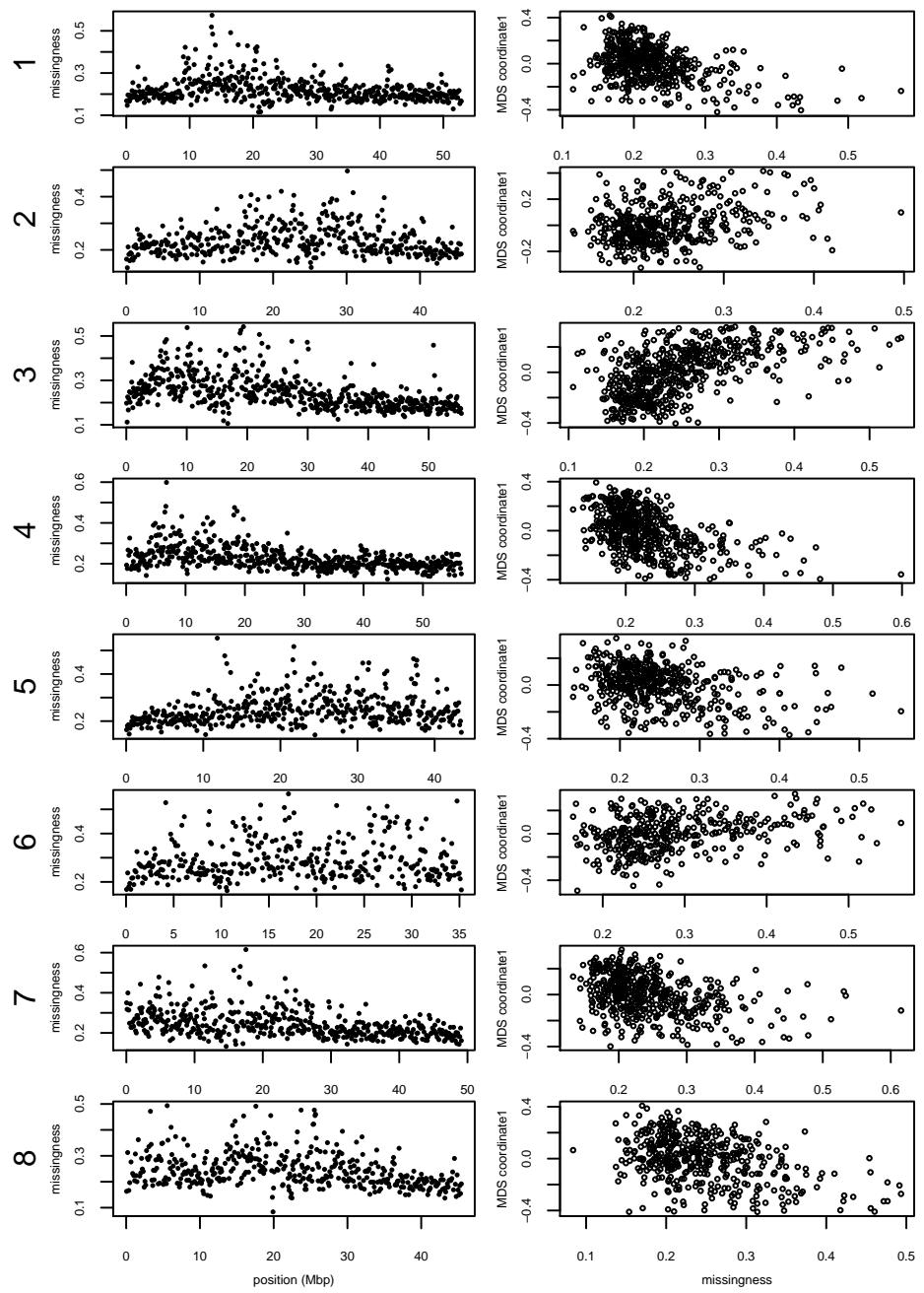


Figure S23: The missingness ratio along genome and the correlation of MDS against missingness in *Medicago*.

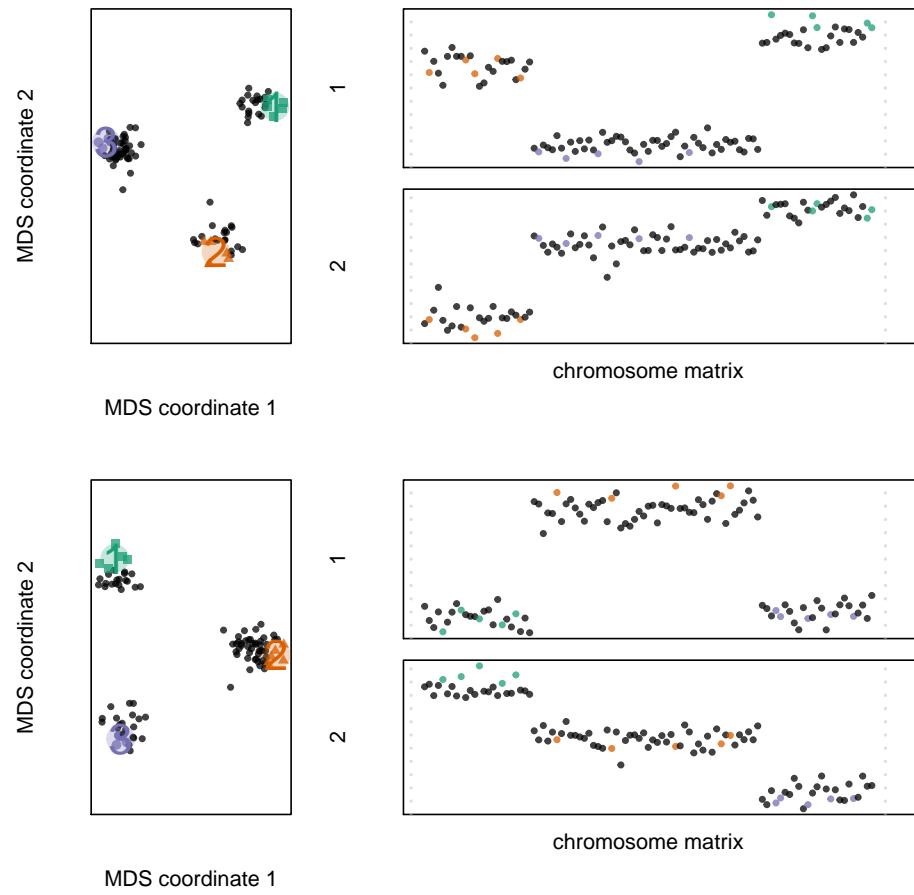


Figure S24: From simple example – base case; and with 50% missing data.

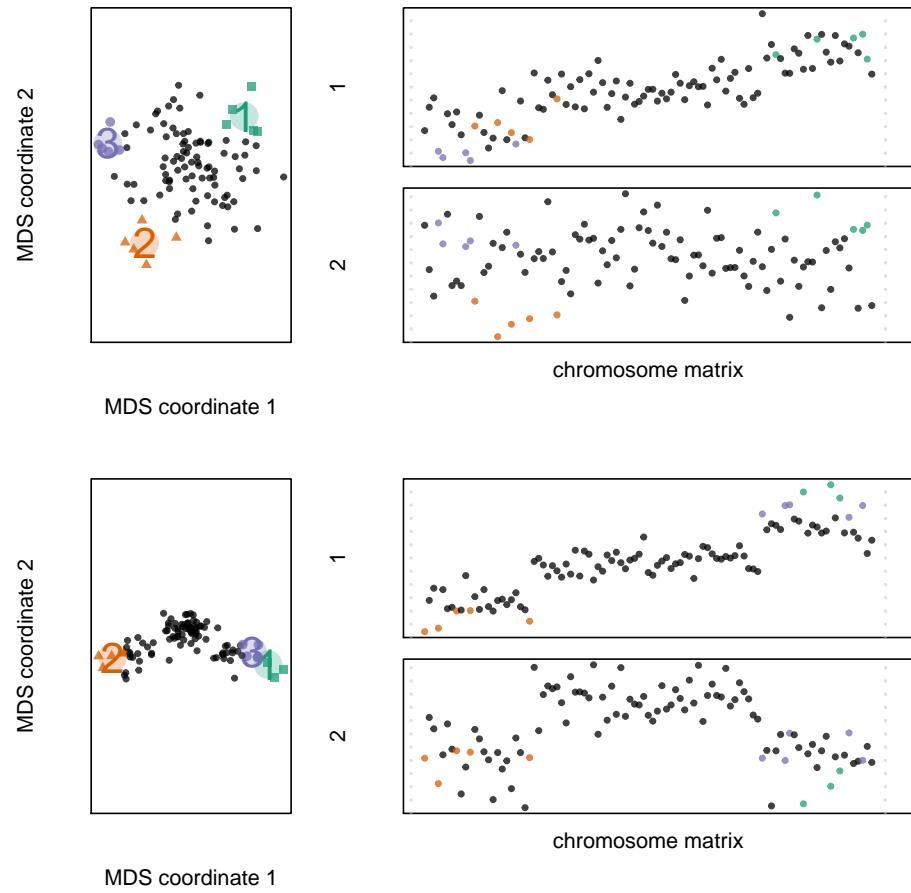


Figure S25: From simple example – different sample sizes; unweighted (top); weighted by $1/\sqrt{n_i}$ (bottom).

Resubmission Cover Letter
Genetics

Han Li
and Peter Ralph
August 21, 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) ... variation in the recombination rate ... creating spurious structure.

Reply: We address this in two ways. This 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, recombination rate variation cannot be driving the results.

(1.2) ... and variation in the mutation rate, creating spurious structure.

Reply:

(1.3) 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. ??, l. ??) and added more explicit notes about this to (p. ??, l. ??) and (p. ??, l. ??).

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

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

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

(1.7) 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 (2008) 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. ??, l. ??).

(1.8) *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 (p. ??, l. ??).

Reviewer 2:

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

(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. ??, l. ??). 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.*

(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.

(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.

(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.

(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.

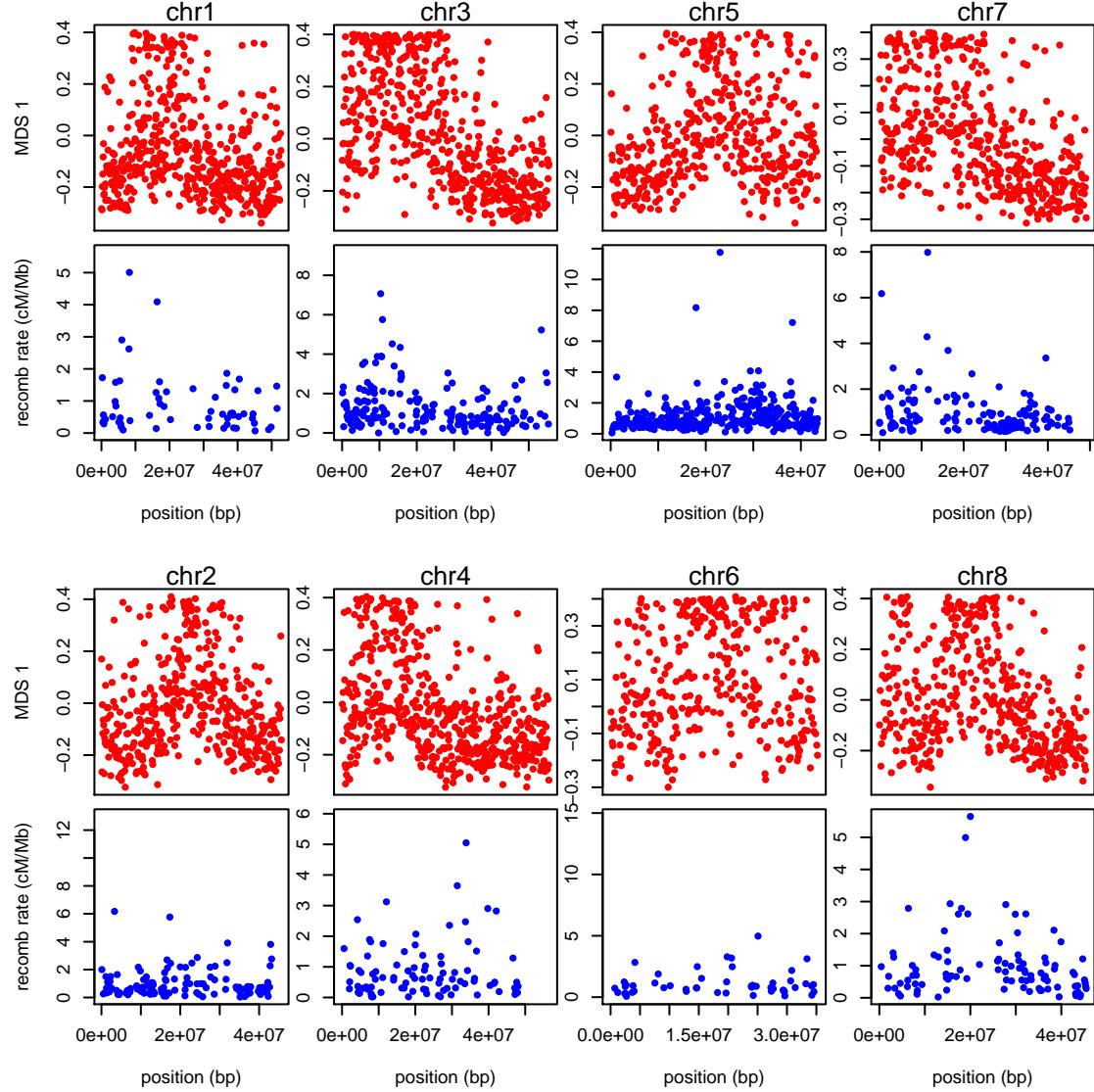
Reply: This is a good point, but XXX We now cite PCAdmix (p. ??, l. ??).

(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.

(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. ??, l. ??)).

(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.

(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“.

(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.

(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?

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

(2.14) Typo in abstract, line 6 “, We show” -*à* “. We show”.

(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”?

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

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

(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 greatly prefer named references rather than numbers, but the L^AT_EXsource code is available at https://github.com/petrelharp/local_pca (run ‘make local_pca_paper.pdf’ in the ‘writeup/’ directory) – the reviewer is welcome to change the formatting and recompile.