

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

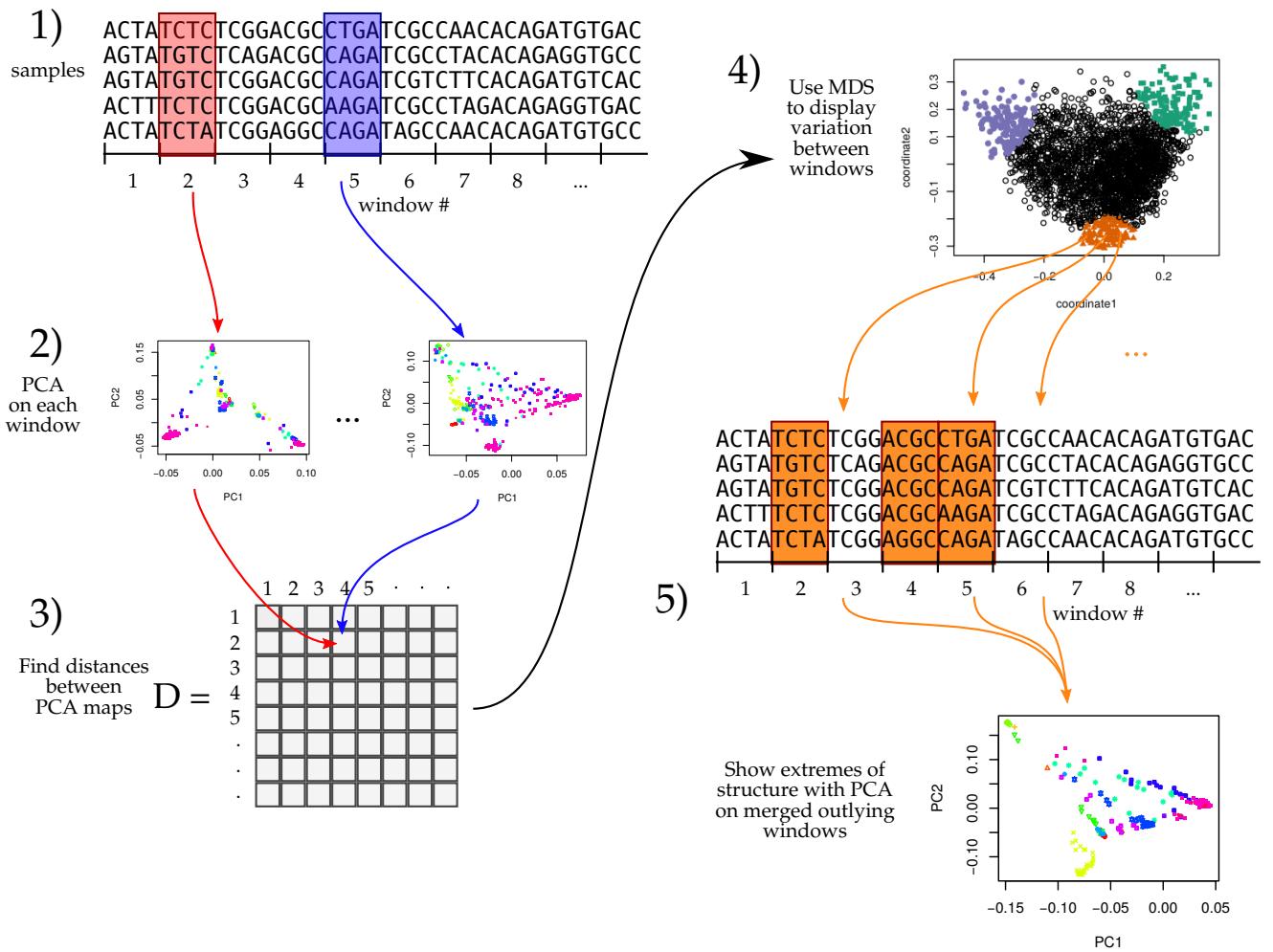


Figure 1: An illustration of the method; see Methods for details.

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 Datasets**

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

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

198 **Human:** We also used genomic data from the entire POPRES dataset (Nelson et al.
199 2008), which has array-derived genotype information for 447,267 SNPs across the 22 au-
200 tosomes of 3,965 samples in total: 346 African-Americans, 73 Asians, 3,187 Europeans
201 and 359 Indian Asians. Since these data derive from genotyping arrays, the SNP density is
202 much lower than the other datasets, which are each derived from whole genome sequencing.
203 We excluded the sex chromosomes and the mitochondria. PCA plots for each chromosome,
204 separately, are shown in Figure S3.

205 ***Medicago truncatula*:** Finally, we used whole-genome sequencing data from the *Med-*
206 *icago truncatula* Hapmap Project (Tang et al. 2014), which has 263 samples from 24 coun-
207 tries, primarily distributed around the Mediterranean basin. Each of the 8 chromosomes
208 has 3–5 million SNPs; PCA plots for these are shown in Figure S4. We did not use the
209 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.

210 2.5 Data access

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

214 Datasets are available as follows: human (POPRES) at dbGaP with accession number
 215 phs000145.v4.p2, *Medicago* at the Medicago Hapmap <http://www.medicagohapmap.org/>,
 216 and *Drosophila* at the Drosophila Genome Nexus, <http://www.johnpool.net/genomes.html>.
 217

218 3 Results

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

225 3.1 Validation

226 Address mutation rate variation; recombination rate variation; choice of number of PCs;
 227 choice of window size; variation in missingness; maybe differing sample sizes and reweighting.
 228

229 3.2 *Drosophila melanogaster*

230 We applied the method to windows of average length 9 Kbp, across chromosome arms
 231 2L, 2R, 3L, 3R and X separately. The first column of Figure 2 is a multidimensional
 232 scaling (MDS) visualization of the matrix of dissimilarities between genomic windows: in
 233 other words, genomic windows that are closer to each other in the MDS plot show more
 234 similar patterns of relatedness. For each chromosome arm, the MDS visualization roughly

235 resembles a triangle, sometimes with additional points. Since the relative position of each
236 window in this plot shows the similarity between windows, this suggests that there are
237 at least three extreme manifestations of population structure typified by windows found
238 in the “corners” of the figure, and that other windows’ patterns of relatedness may be a
239 mixture of those extremes. The next two columns of Figure 2 respectively depict the two
240 MDS coordinates of each window, plotted against the window’s position along the genome,
241 to show how the plot of the first column is laid out along the genome.

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

250 The most striking variation in patterns of relatedness turns out to be explained by
251 several large inversions that are polymorphic in these samples, discussed in Corbett-Detig
252 and Hartl (2012) and Langley et al. (2012). To depict this, Figure 3 shows the PCA plots in
253 Figure S1 recolored by the orientation of the inversion for each sample. Taking chromosome
254 arm 2L as an example, the two regions of similar, extreme patterns of relatedness shown
255 in green in the first row of Figure 2 lie directly around the breakpoints of the inversion
256 In(2L)t, and the PCA plots in the first rows of Figure 3 shows that patterns of relatedness
257 here are mostly determined by inversion orientation. The regions shown in purple on
258 chromosome 2L lie near the centromere, and have patterns of relatedness reflective of two
259 axes of variation, seen in Figures S1 and 3, which correspond roughly to latitude within
260 Africa and to degree of cosmopolitan admixture respectively (see Lack et al. (2015) for more
261 about admixture in this sample). The regions shown in orange on chromosome 2L mostly lie
262 inside the inversion, and show patterns of relatedness that are a mixture between the other
263 two, as expected due to recombination within the (long) inversion (Guerrero et al. 2011).
264 Similar results are found in other chromosome arms, albeit complicated by the coexistence
265 of more than one polymorphic inversion; however, each breakpoint visibly affects patterns
266 in the MDS coordinates (see vertical lines in Figure 2).

267 To see how patterns of relatedness vary in the absence of polymorphic inversions, we
268 performed the same analyses after removing, for each chromosome arm, any samples car-
269 rying inversions on that arm. In the result, shown in Supplemental Figure S5, the striking
270 peaks associated with inversion breakpoints are gone, and previously smaller-scale vari-
271 ation now dominates the MDS visualization. For instance, the majority of the variation
272 along 3L in Figure 2 is on the left end of the arm, dominated by two large peaks around the
273 inversion breakpoints; there is also a relatively small dip on the right end of the arm (near
274 the centromere). In contrast, Supplemental Figure S5 shows that after removing polymor-
275 phic inversions, remaining structure is dominated by the dip near the centromere. Without

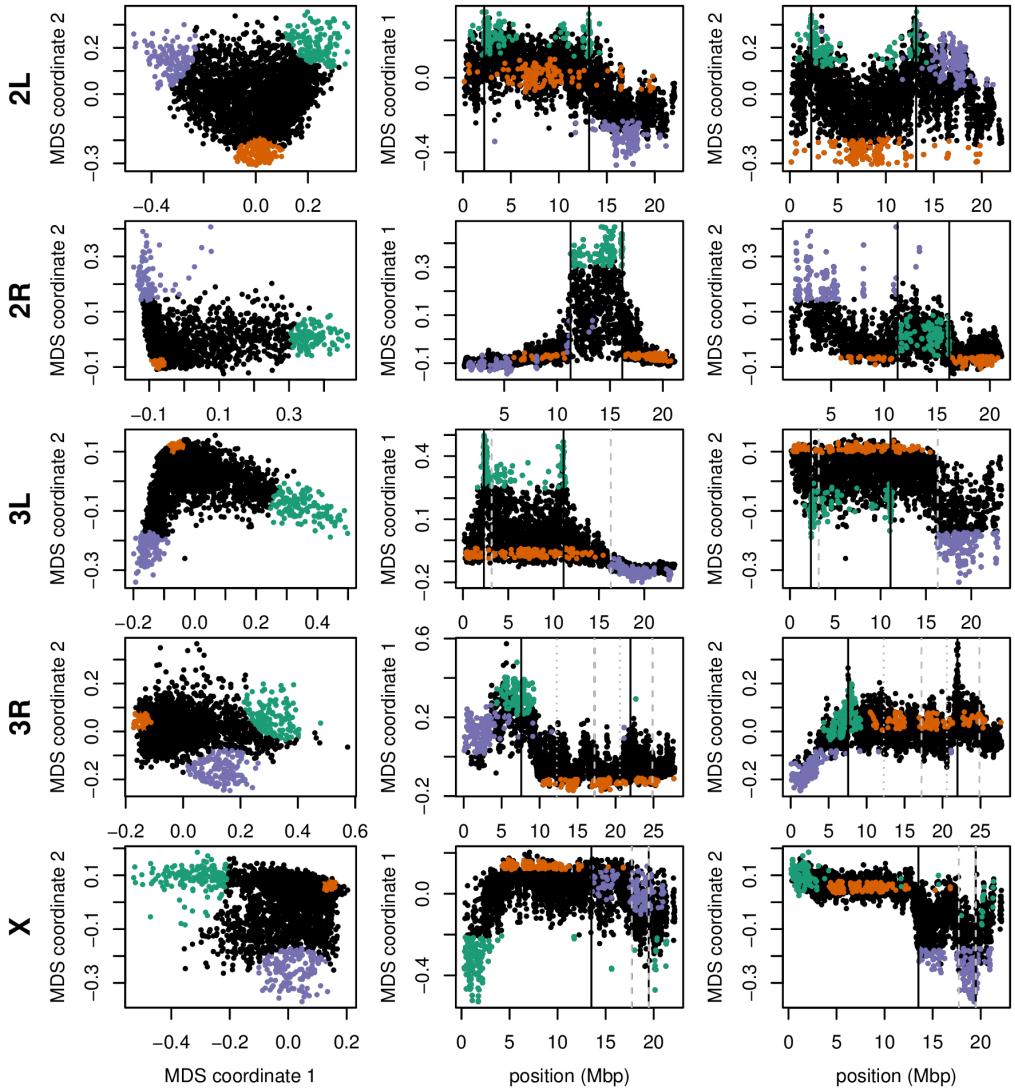


Figure 2: Variation in patterns of relatedness for windows across *Drosophila melanogaster* chromosome arms. In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the two MDS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

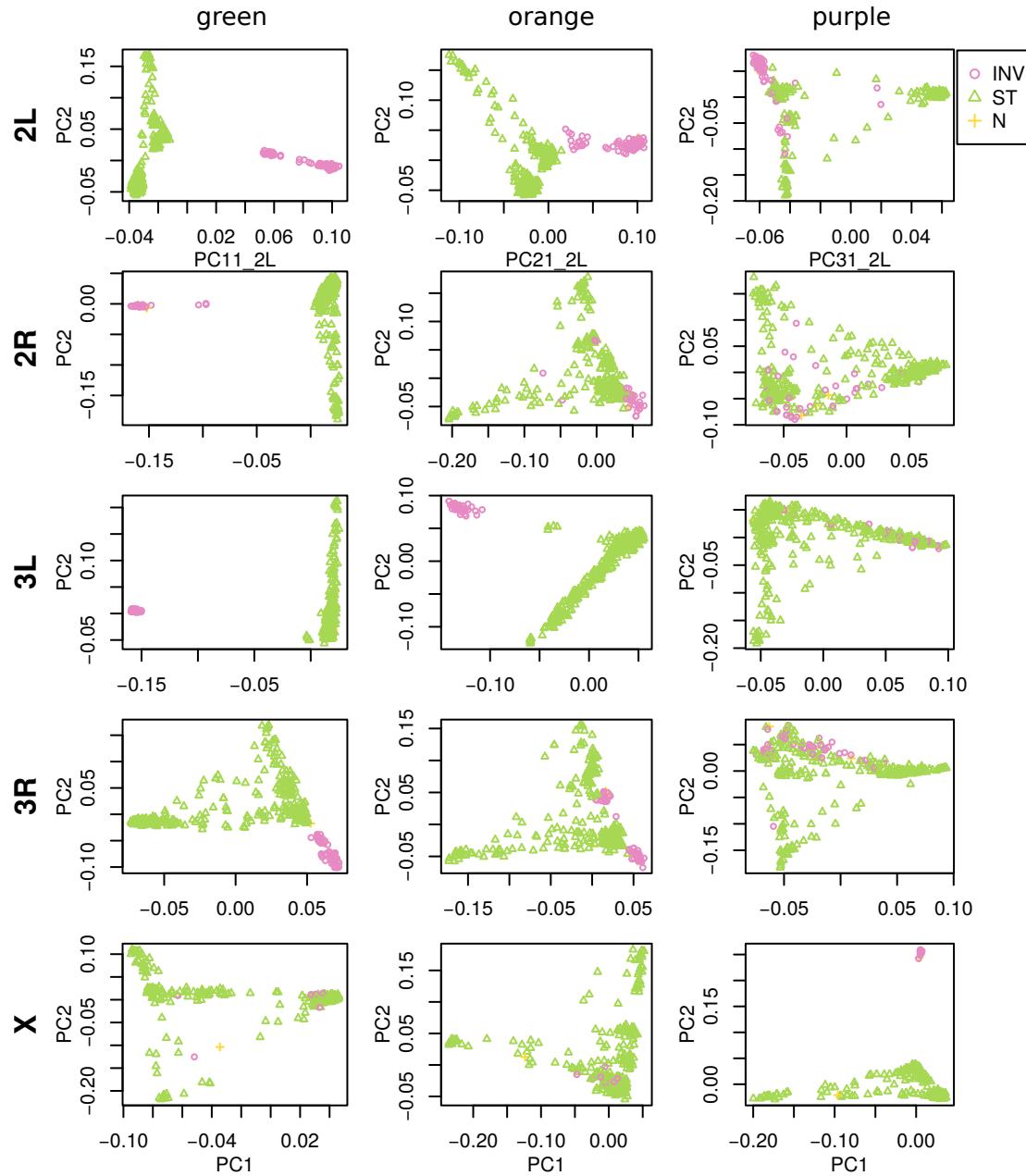


Figure 3: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange windows; and third is for purple windows. In each, samples are colored by orientation of the polymorphic inversions In(2L)t, In(2R)NS, In(3L)OK, In(3R)K and In(1)A respectively (data from (Lack et al. 2015)). In each “INV” denotes an inverted genotype, “ST” denotes the standard orientation, and “N” denotes unknown.

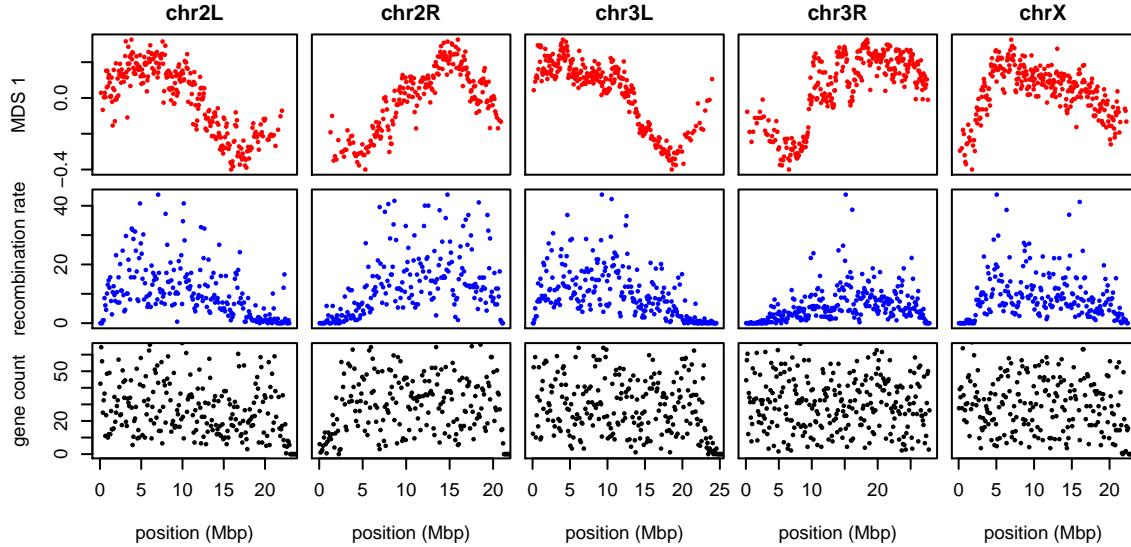


Figure 4: The effects of population structure without inversions is correlated to recombination rate in *Drosophila melanogaster*. The first plot (in red) shows the first MDS coordinate along the genome for windows of 10,000 SNPs, obtained after removing samples with inversions. (A plot analogous to 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.

276 inversions, variation in patterns of relatedness shown in the MDS plots follows similar pat-
 277 terns to that previously seen in *D. melanogaster* recombination rate and diversity (Langley
 278 et al. 2012; Mackay et al. 2012). Indeed, correlations between the recombination rate in
 279 each window and the position on the first MDS coordinate are highly significant (Spear-
 280 man's $\rho = 0.54$, $p < 2 \times 10^{-16}$; Figures 4 and S6). This is consistent with the hypothesis
 281 that variation is due to selection, since the strength of linked selection increases with local
 282 gene density, measured in units of recombination distance. The number of genes – mea-
 283 sured as the number of transcription start and end sites within each window – was not
 284 significantly correlated with MDS coordinate ($p = 0.22$).

285 **3.3 Human**

286 As we did for the *Drosophila* data, we applied our method separately to all 22 human
287 autosomes. On each, variation in patterns of relatedness was dominated by a small number
288 of windows having similar patterns of relatedness to each other that differed dramatically
289 from the rest of the chromosome. These may be primarily inversions: outlying windows
290 coincide with three of the six large polymorphic inversions described in Antonacci et al.
291 (2009), notably a particularly large, polymorphic inversion on 8p23 (Figure 5). Similar
292 plots for all chromosomes are shown in Supplementary Figures S7, S8, and S9. PCA plots
293 of many outlying windows show a characteristic trimodal shape (shown for chromosome 8 in
294 Figure S10), presumably distinguishing samples having each of the three diploid genotypes
295 for each inversion orientation (although we do not have data on orientation status). This
296 trimodal shape has been proposed as a method to identify inversions (Ma and Amos 2012),
297 but distinguishing this hypothesis from others, such as regions of low recombination rate,
298 would require additional data.

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

303 **3.4 *Medicago truncatula***

304 Unlike the other two species, the method applied separately on all eight chromosomes of
305 *Medicago truncatula* showed similar patterns of gradual change in patterns of relatedness
306 across each chromosome, with no indications of chromosome-specific patterns. This con-
307 sistency suggests that the factor affecting the population structure for each chromosome
308 is the same, as might be caused by varying strengths of linked selection. To verify that
309 variation in the effects of population structure is shared across chromosomes, we applied
310 the method to all chromosomes together. Results for chromosome 3 are shown in Figures
311 6 and 6, and other chromosomes are similar: across chromosomes, the high values of the
312 first MDS coordinate coincide with the position of the heterochromatic regions surrounding
313 the centromere, which often have lower gene density and may therefore be less subject to
314 linked selection. To verify that this is a possible explanation, we counted the number of
315 genes found in each window using gene models in Mt4.0 from jcvi.org (Tang et al. 2014),
316 which are shown juxtaposed with the first MDS coordinate of each window in Figure 7,
317 and are significantly correlated, as shown in Supplemental Figure S12. (Values shown are
318 the number of start and end positions of each predicted mRNA transcript, divided by two,
319 assigned to the nearest window.) However, other genomic features, such as distance to
320 centromere show roughly the same patterns, so we cannot rule out alternative hypotheses.
321 In particular, fine-scale recombination rate estimates are not available in a form mappable
322 to Mt4.0 coordinates (although those in Paape et al. (2012) appear visually similar).

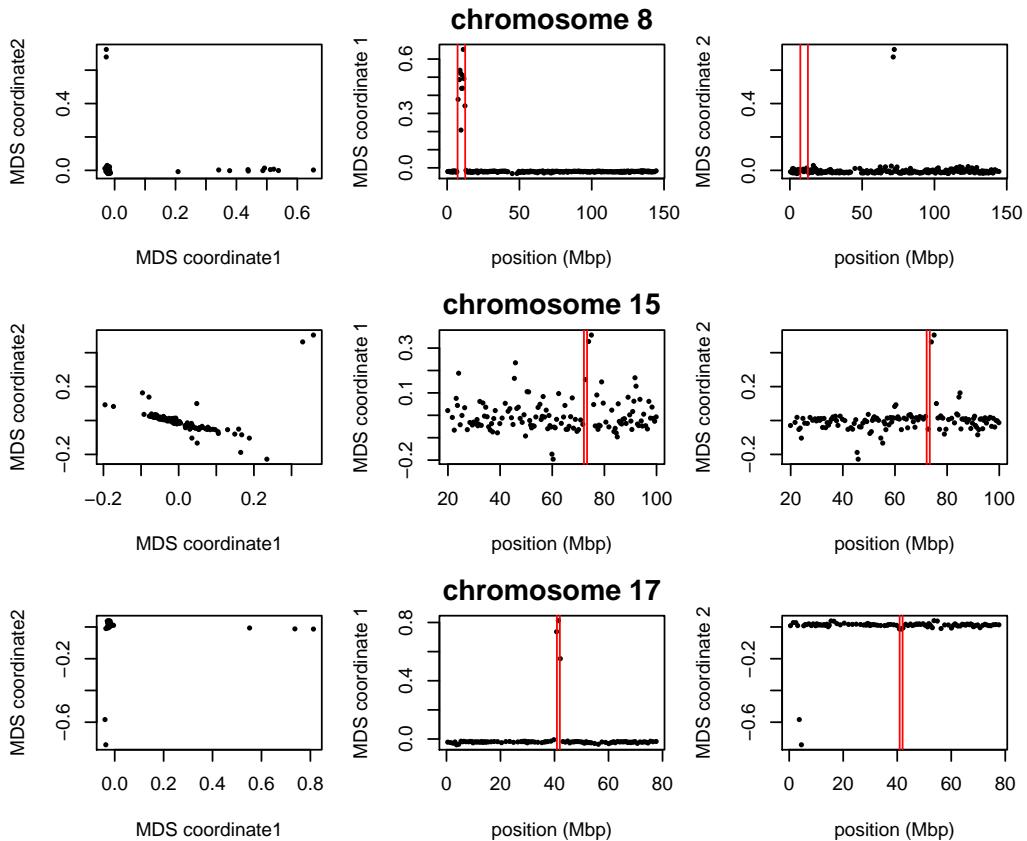


Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in each plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position (midpoint) along the chromosome. Rows, from top to bottom show chromosomes 8, 15, and 17. The vertical red lines show the breakpoints of known inversions from Antonacci et al. (2009).

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

327 4 Discussion

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

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

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

355 **The effect of selection** It seems that the variation in patterns of relatedness we see in
356 the *Medicago truncatula* and *Drosophila melanogaster* datasets must be explained some-
357 how by linked selection. Furthermore, the selection must be affecting many targets across
358 the genome, since we see similar effects across long distances (even distinct chromosomes).

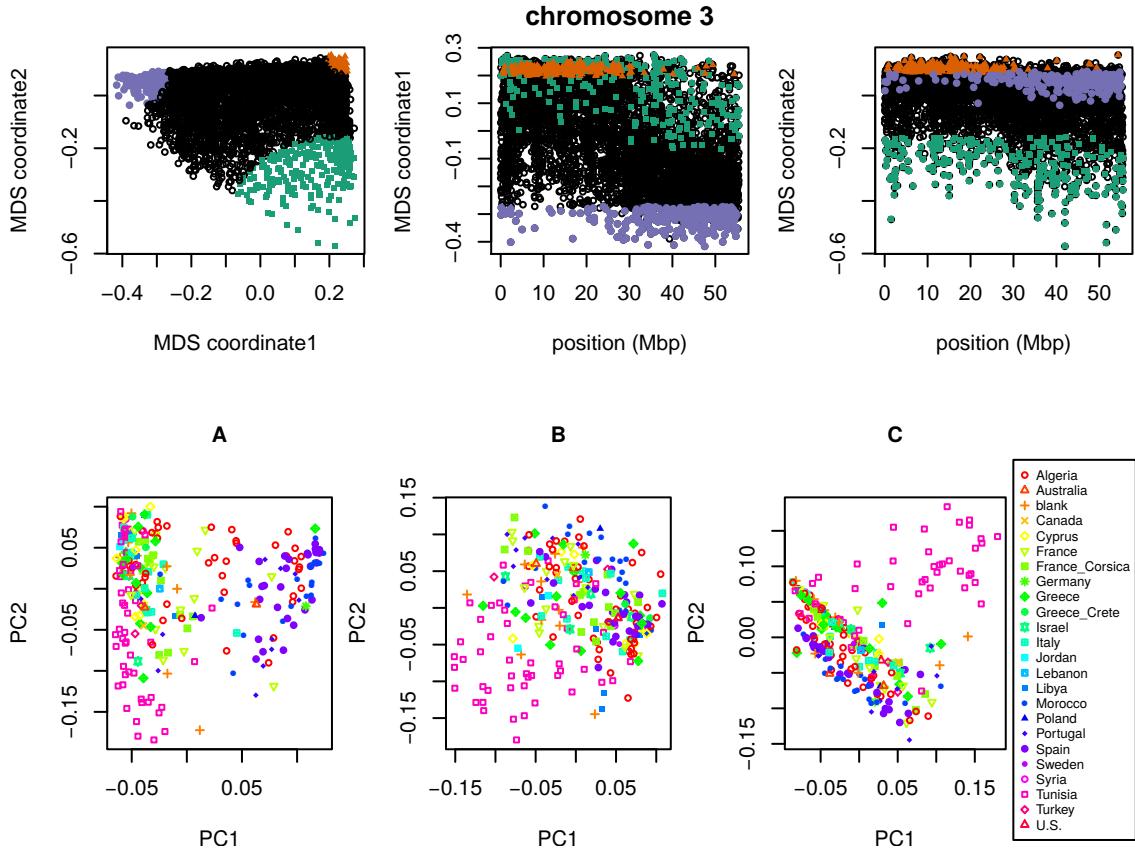


Figure 6: MDS visualization of patterns of relatedness on *M. truncatula* chromosome 3, with corresponding PCA plots. Each point in the plot represents a window; the structure revealed by the MDS plot is strongly clustered along the chromosome, with windows in the upper-right corner of the MDS plot (colored 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.

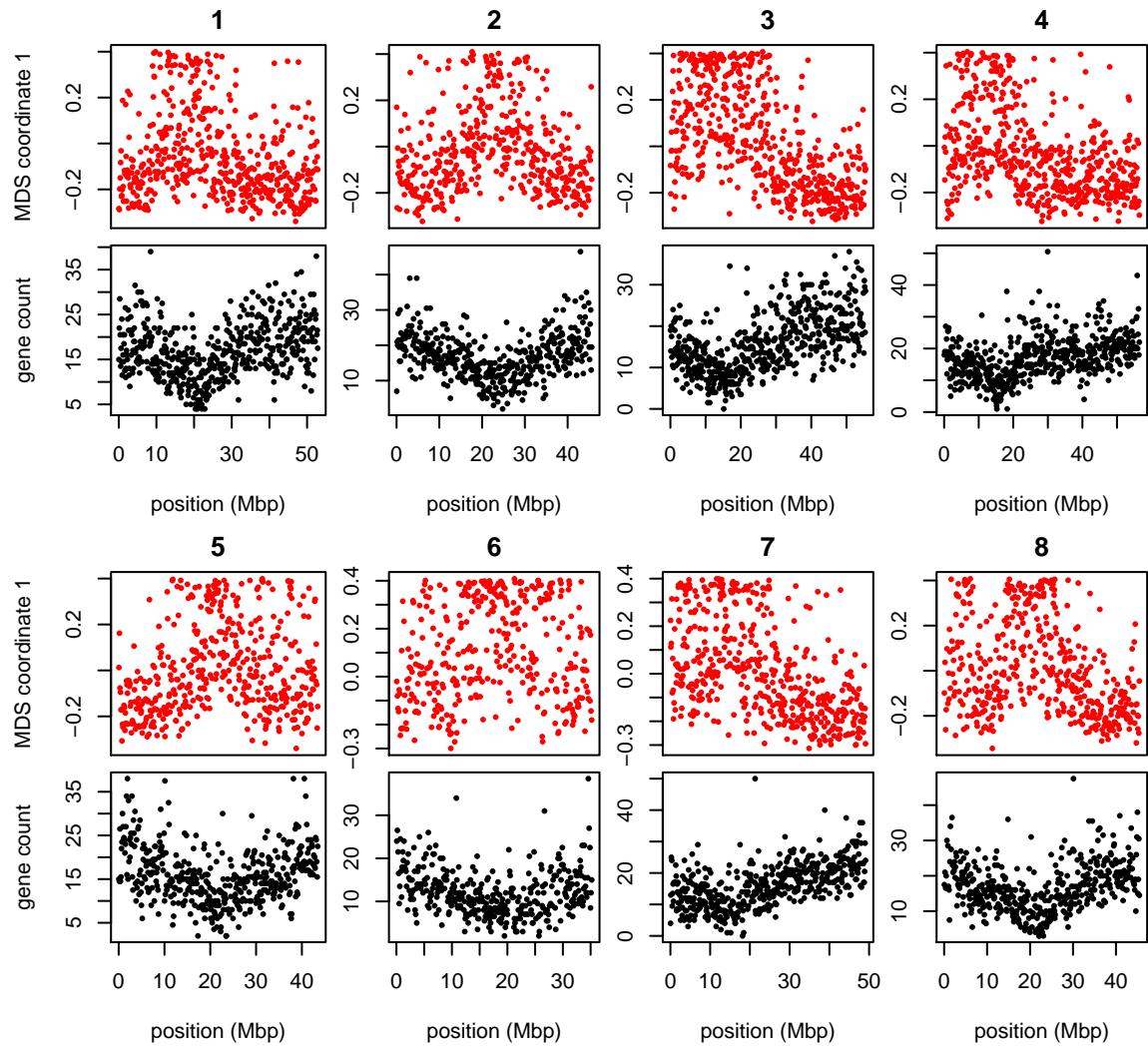


Figure 7: MDS coordinate and gene density for each window in the *Medicago* genome, for chromosomes 1–8 (numbered above each pair of figures). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene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

359 For this reason, the most likely candidate may be selection against linked deleterious mu-
360 tations, known as “background selection” (Charlesworth et al. 1993; Charlesworth 2013).
361 Informally, background selection reduces the number of potential contributors to the gene
362 pool in regions of the genome with many possible deleterious mutations (Hudson and Ka-
363 plan 1995); for this reason, if it acts in a spatial context, it is expected to induce samples
364 from nearby locations to cluster together more frequently. Therefore, regions of the genome
365 harboring many targets of local adaptation may show similar patterns, since migrant alleles
366 in these regions will be selected against, and so locally gene trees will more closely reflect
367 spatial proximity.

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

377 There has been substantial debate over the relative impacts of different forms of se-
378 lection. These have been difficult to disentangle in part because for the most part theory
379 makes predictions which are only strictly valid in randomly mating (i.e., unstructured)
380 populations, and it is unclear to what extent the spatial structure observed in most real
381 populations will affect these predictions. It may be possible to design more powerful statis-
382 tics that make stronger use of spatial information.

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

396 Besides window length, there is also the question of how to choose windows. In these
397 applications we have used nonoverlapping windows with equal numbers of polymorphic
398 sites. Alternatively, windows could be chosen to have equal length in genetic distance, so

399 that each would have roughly the same number of independent trees. However, we found
400 little change in results when using different window sizes or when measuring windows in
401 physical distance (in bp).

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

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

420 Acknowledgements

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

427 Disclosure declaration

428 The authors declare no conflicts of interest.

429 A Choosing window length

430 The choice of window length entails a balance between signal and noise. In very short
431 windows, genealogies of the samples will only be represented by a few trees, so varia-

432 tion between windows represents demographic noise rather than meaningful variation in
 433 patterns of relatedness. Longer windows generally have more distinct trees (and SNPs), al-
 434 lowing for less noisy estimation of local patterns of relatedness. However, to better resolve
 435 meaningful signal, i.e., differences in patterns of relatedness along the genome, we would
 436 like reasonably short windows.

437 Since we summarize patterns of relatedness using relative positions in the principal
 438 component maps, we quantify “noise” as the standard error of a sample’s position on PC1
 439 in a particular window, averaged across windows and samples, and “signal” as the standard
 440 deviation of the sample’s position on PC1 over all windows, averaged over samples. The
 441 definition of eigenvectors does not specify their sign, and so when comparing between
 442 windows we choose signs to best match each other: after choosing $PC1_1$, for instance,
 443 if u is the first eigenvector obtained from the covariance matrix for window j , then we
 444 next choose $PC1_j = \pm u$, where the sign is chosen according to which of $\|PC1_1 - u\|$ or
 445 $\|PC1_1 + u\|$ is smaller.

446 After doing this, the mean variance across windows is

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

447 where $PC1_{ij}$ is the position of the i^{th} individual on $PC1$ in window j , and $\overline{PC1}_j =$
 448 $(1/N) \sum_{j=1}^N PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jack-
 449 knife (Busing et al. 1999; Efron 1982): we divide the j^{th} window into 10 equal-sized
 450 pieces, and let $PC1_{ij,k}$ denote the first principal component of this region found af-
 451 ter removing the k^{th} piece; then the estimate of the squared standard error is $\sigma_{ij}^2 =$
 452 $\frac{9}{10} \sum_{k=1}^{10} (PC1_{ij,k} - \frac{1}{10} \sum_{\ell=1}^{10} PC1_{ij,\ell})^2$. Averaging over samples and windows,

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

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

460 B Weighted PCA

461 Principal components analysis can be thought of as finding a good low-dimensional matrix
 462 factorization (Engelhardt and Stephens 2010) that well-approximates the original data in

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

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

463 the least-squares sense: if C is the $N \times N$ genetic covariance matrix, then to find the top k
 464 principal components, we find an orthogonal $N \times k$ matrix U , and a $k \times k$ diagonal matrix
 465 Λ with diagonal entries $\Lambda_{ii} = \lambda_i$ to minimize

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

466 The columns of U , known as the principal components, are the eigenvectors of C , the
 467 entries of λ are the eigenvalues of C , and the proportion of variance explained by the m^{th}
 468 component is

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

469 Thinking about the problem as a least-squares approximation problem makes it clear
 470 why unbalanced sample sizes can result in undesirable outcomes. If we want to describe
 471 variation *between* populations, but 80% of the samples are from a single population, then
 472 unless populations are highly differentiated, a better approximation to C may be obtained
 473 by using the columns of U to describe variation *within* the overrepresented population
 474 rather than between the populations. A common workaround is to remove samples, but a
 475 more elegant solution can be found by reweighting the objective function in (3). Let w_i be

⁴⁷⁶ a weight associated with sample i , W the diagonal matrix with w along the diagonal, and
⁴⁷⁷ instead seek to minimize

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

⁴⁷⁸ and now for convenience we require U to be orthogonal in $\ell_2(w)$, i.e., that $U^T W U = I$.
⁴⁷⁹ We then would choose w to give roughly equal weight to each *population*, instead of each
⁴⁸⁰ individual. We have used with good results the weightings $w_i = 1/\max(10, n_i)$, where
⁴⁸¹ n_i is, if there are discrete populations, the number of samples in the same population as
⁴⁸² sample i ; or, for continuously sampled individuals, the number of samples within a certain
⁴⁸³ distance of sample i .

⁴⁸⁴ To solve (4), let λ and V denote the top k eigenvalues and eigenvectors of $W^{1/2} C W^{1/2}$,
⁴⁸⁵ so that $V \Lambda V^T$ is the rank k matrix closest in least squares to $W^{1/2} C W^{1/2}$; so if we define
⁴⁸⁶ $U = W^{-1/2} V$ then $U^T W U = V^T V = I$, and

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

⁴⁸⁷ is the low-dimensional approximation to C . The proportion of variance explained is calculated from eigenvalues as before, but has the interpretation

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

⁴⁸⁹ In our R implementation we use the Spectra library (Qiu and Mei 2016) to find only the
⁴⁹⁰ top k eigenvectors.

⁴⁹¹ C Supplementary Tables

	10000snp, 2 PCs	1000snp, 2 PCs	10000snp, 5 PCs	100000bp, 2 PCs	10000bp, 2 PCs
10000snp, 2 PCs	1.00	0.87	0.96	0.90	0.
1000snp, 2 PCs	0.68	1.00	0.73	0.68	0.
10000snp, 5 PCs	0.96	0.92	1.00	0.88	0.
100000bp, 2 PCs	0.90	0.87	0.88	1.00	0.
10000bp, 2 PCs	0.68	0.93	0.72	0.67	1.

⁴⁹² [[2]]

⁴⁹³ D Supplementary Figures

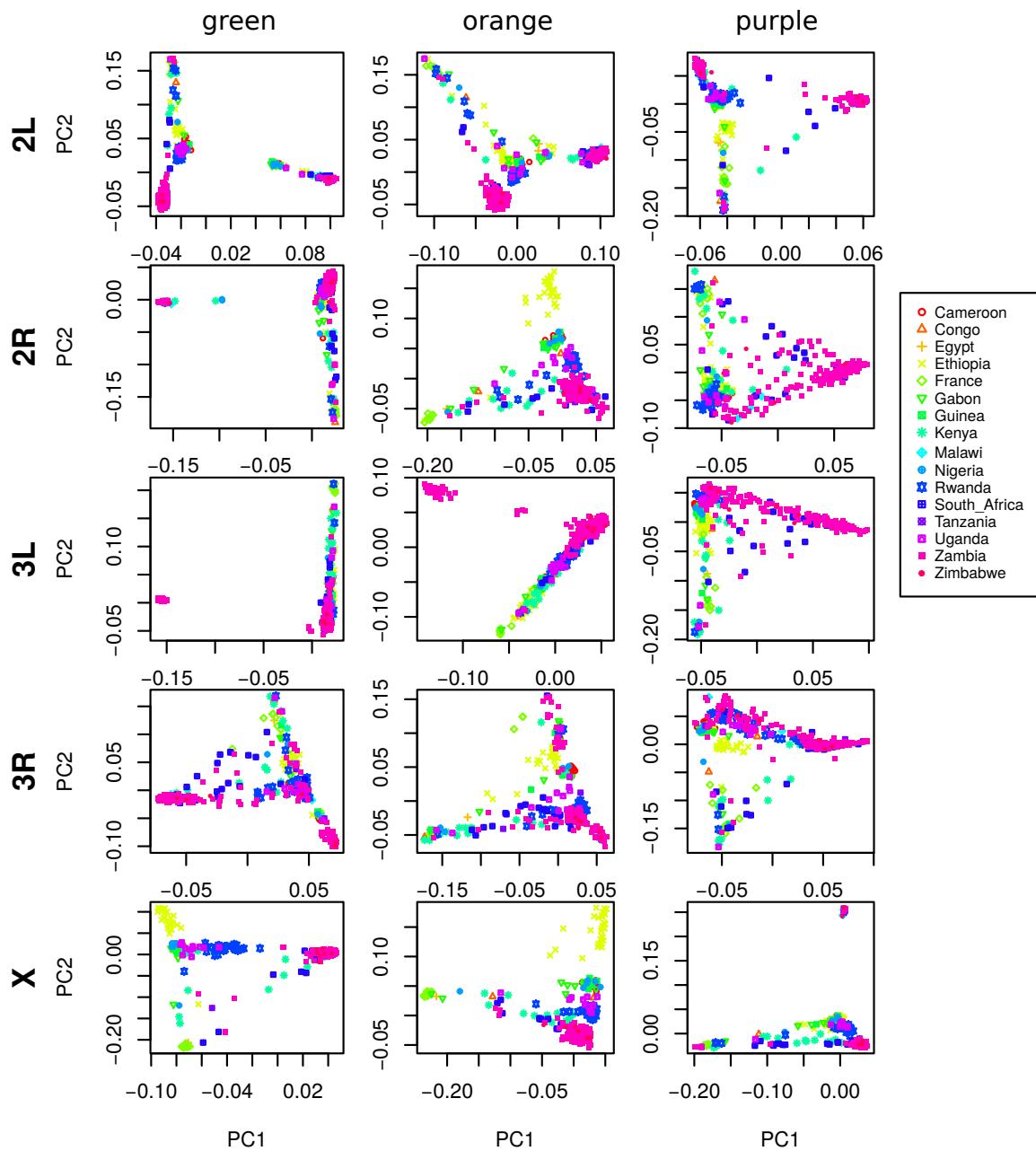


Figure S1: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange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

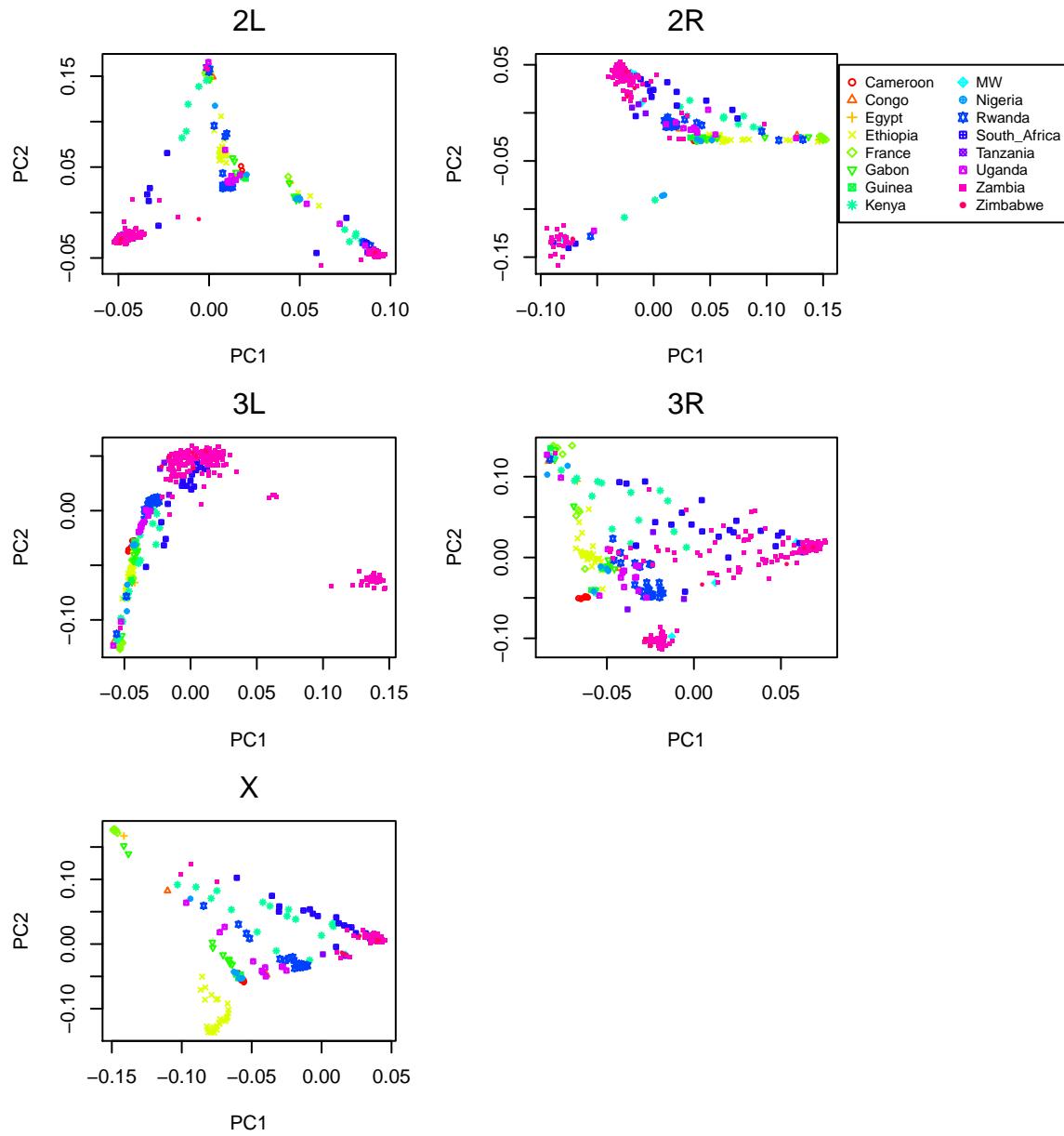


Figure S2: PCA plots for chromosome arms 2L, 2R, 3L, 3R and X of the *Drosophila melanogaster* dataset.

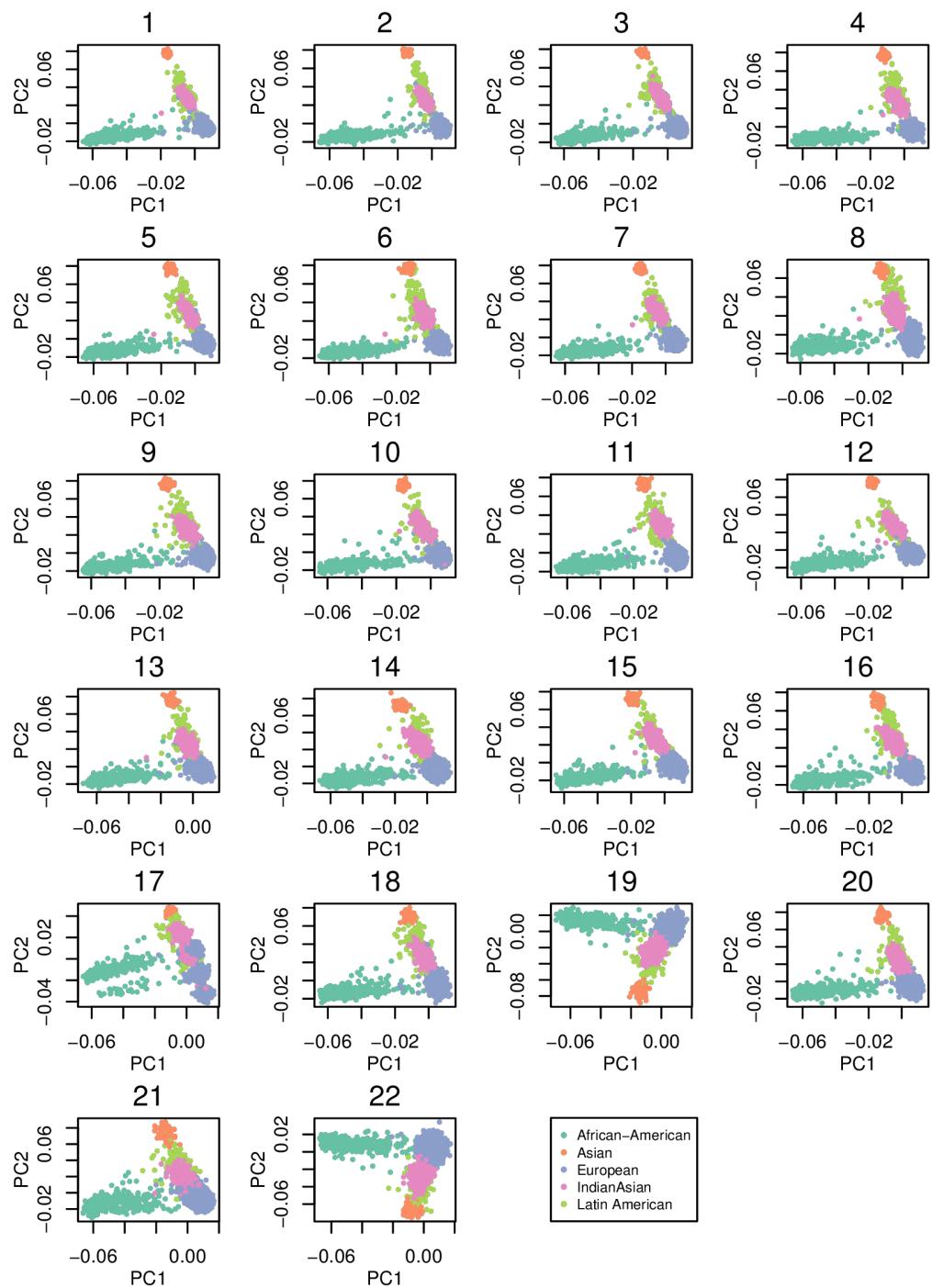


Figure S3: PCA plots for all 22 human autosomes from the POPRES data.

MDS1	10000 SNPs, 2 PCs	1000 SNPs, 2 PCs	10000 SNPs, 5 PCs	100000bp, 2 PCs	100000bp, 2 PCs
10000 SNPs, 2 PCs	1.00	0.54	0.93	0.87	0.87
1000 SNPs, 2 PCs	0.82	1.00	0.76	0.83	0.83
10000 SNPs, 5 PCs	0.93	0.50	1.00	0.83	0.83
100000bp, 2 PCs	0.87	0.59	0.84	1.00	0.84
10000bp, 2 PCs	0.83	0.92	0.77	0.84	0.84
MDS2	10000snp, 2 PCs	1000snp, 2 PCs	10000snp, 5 PCs	100000bp, 2 PCs	10000bp, 2 PCs
10000snp, 2 PCs	1.00	0.54	0.93	0.87	0.87
1000snp, 2 PCs	0.82	1.00	0.76	0.83	0.83
10000snp, 5 PCs	0.93	0.50	1.00	0.83	0.83
100000bp, 2 PCs	0.87	0.59	0.84	1.00	0.84
10000bp, 2 PCs	0.83	0.92	0.77	0.84	0.84

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](#).

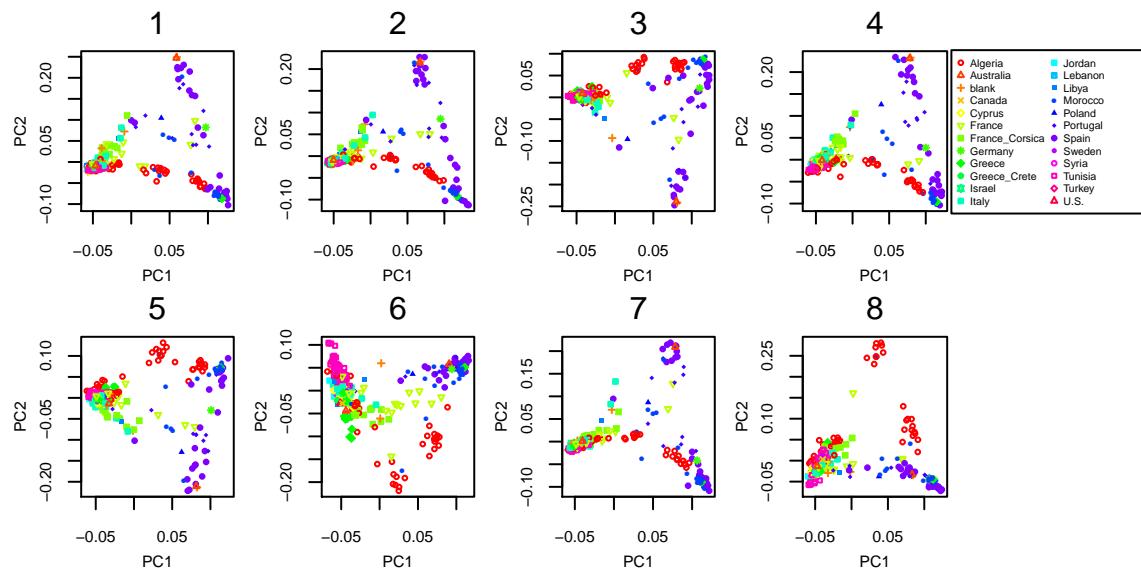


Figure S4: PCA plots for all 8 chromosomes in the *Medicago truncatula* dataset.

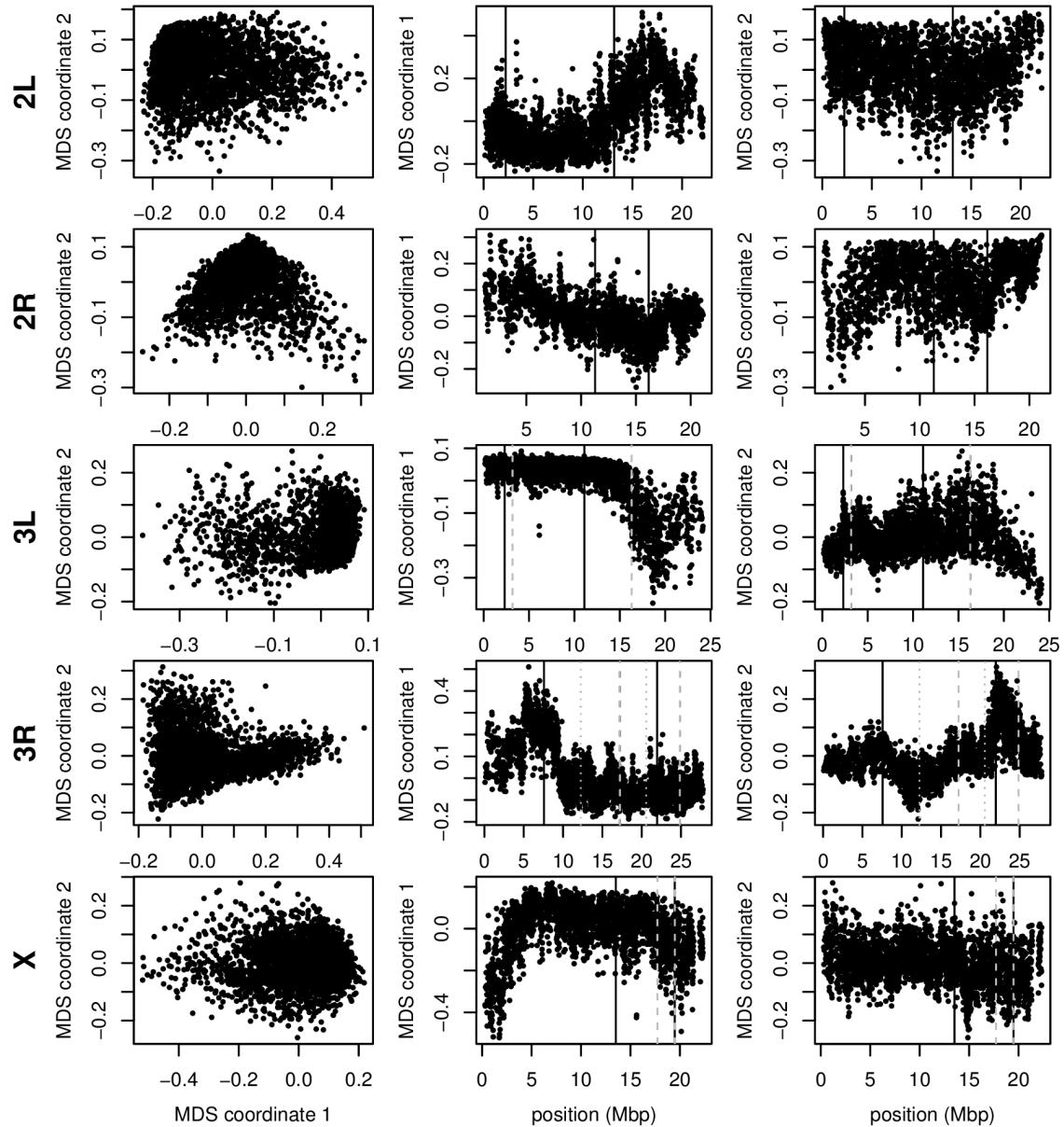


Figure S5: Variation in structure for windows of 1,000 SNPs across *Drosophila melanogaster* chromosome arms: without inversions. As in Figure 2, but after omitting for each chromosome arm individuals carrying the less frequent orientation of any inversions on that chromosome arm. The values differ from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

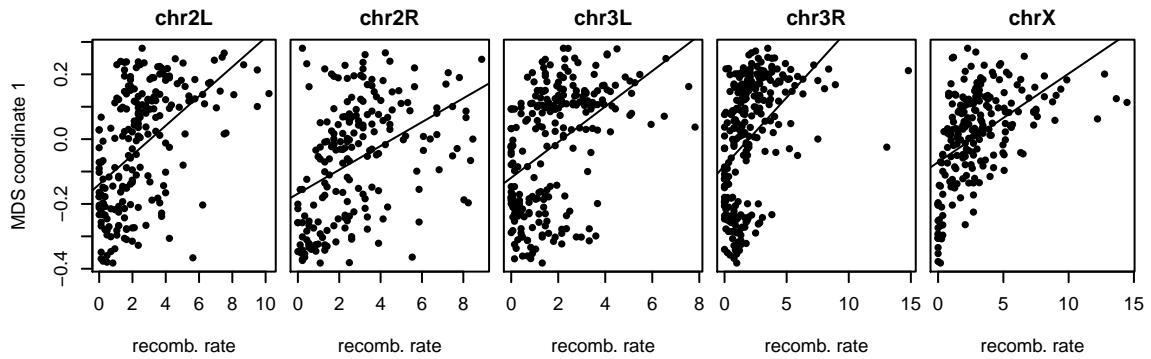


Figure S6: Recombination rate, and the effects of population structure for *Drosophila melanogaster*: this shows the first MDS coordinate and recombination rate (in cM/Mbp), as in Figure 4, against each other. Since the windows underlying estimates of Figure 4 do not coincide, to obtain correlations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

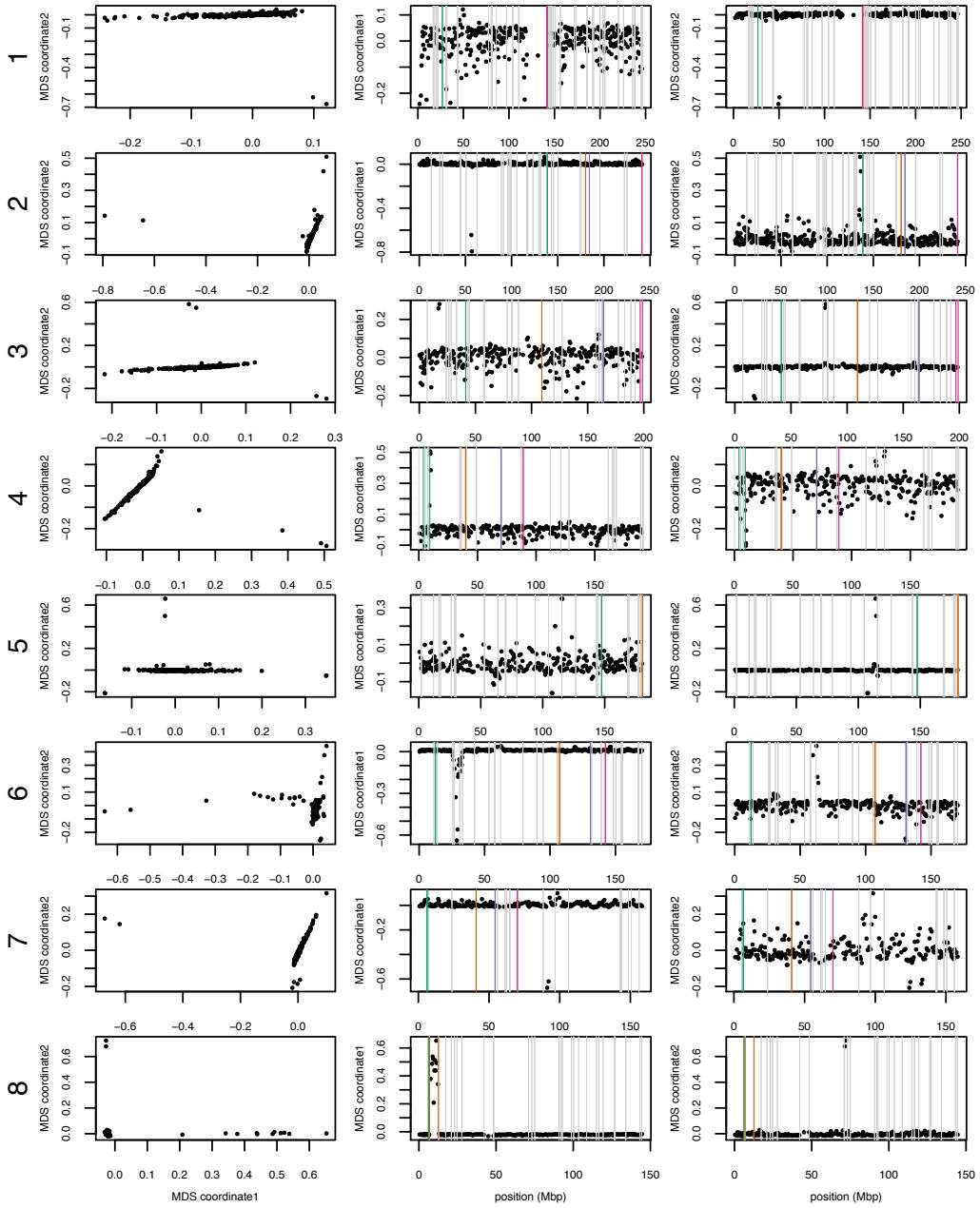


Figure S7: MDS plots for human chromosomes 1-8. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosomes. Colorful vertical lines show the breakpoints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

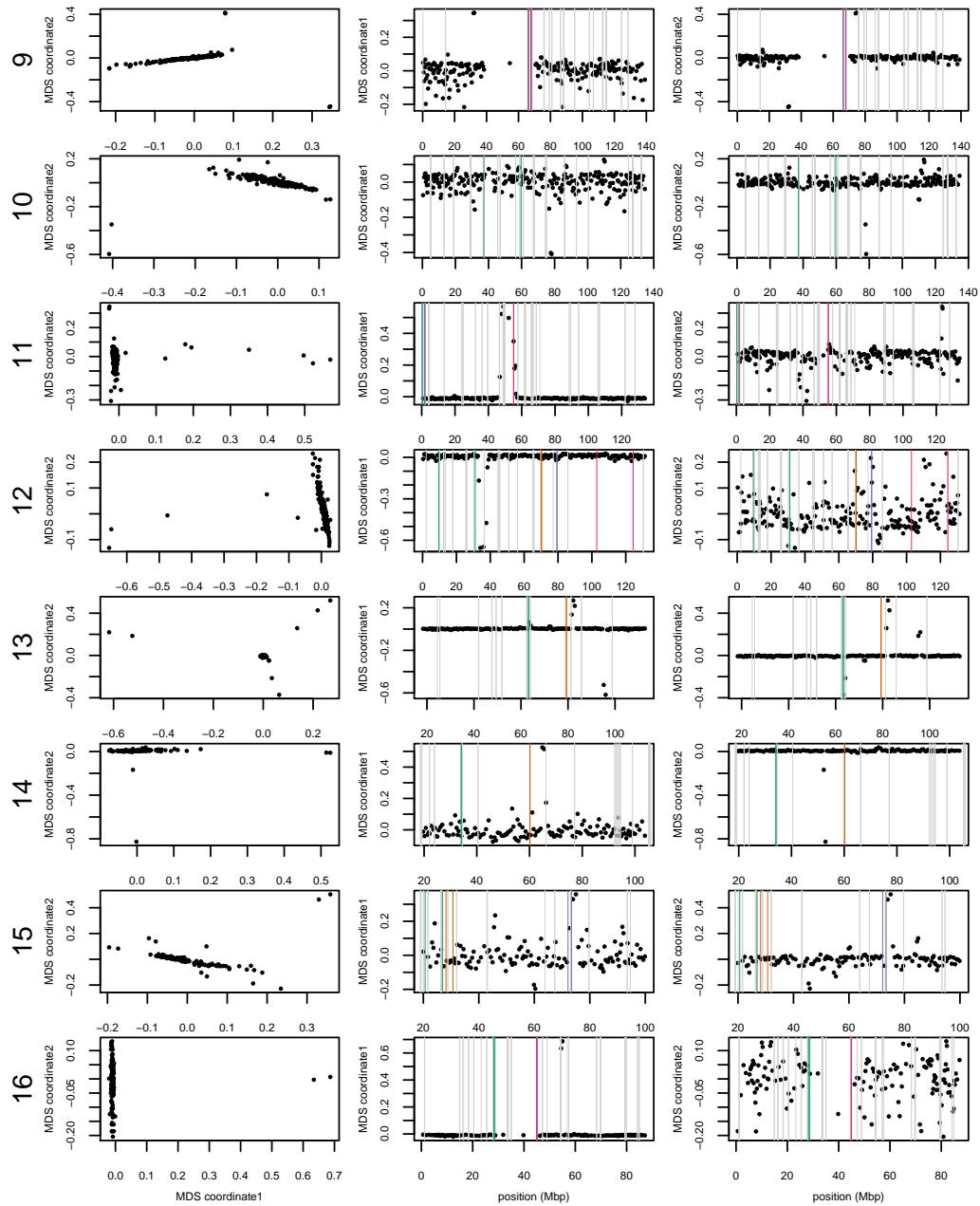


Figure S8: MDS plots for human chromosomes 9-16, as in Supplemental Figure S7.

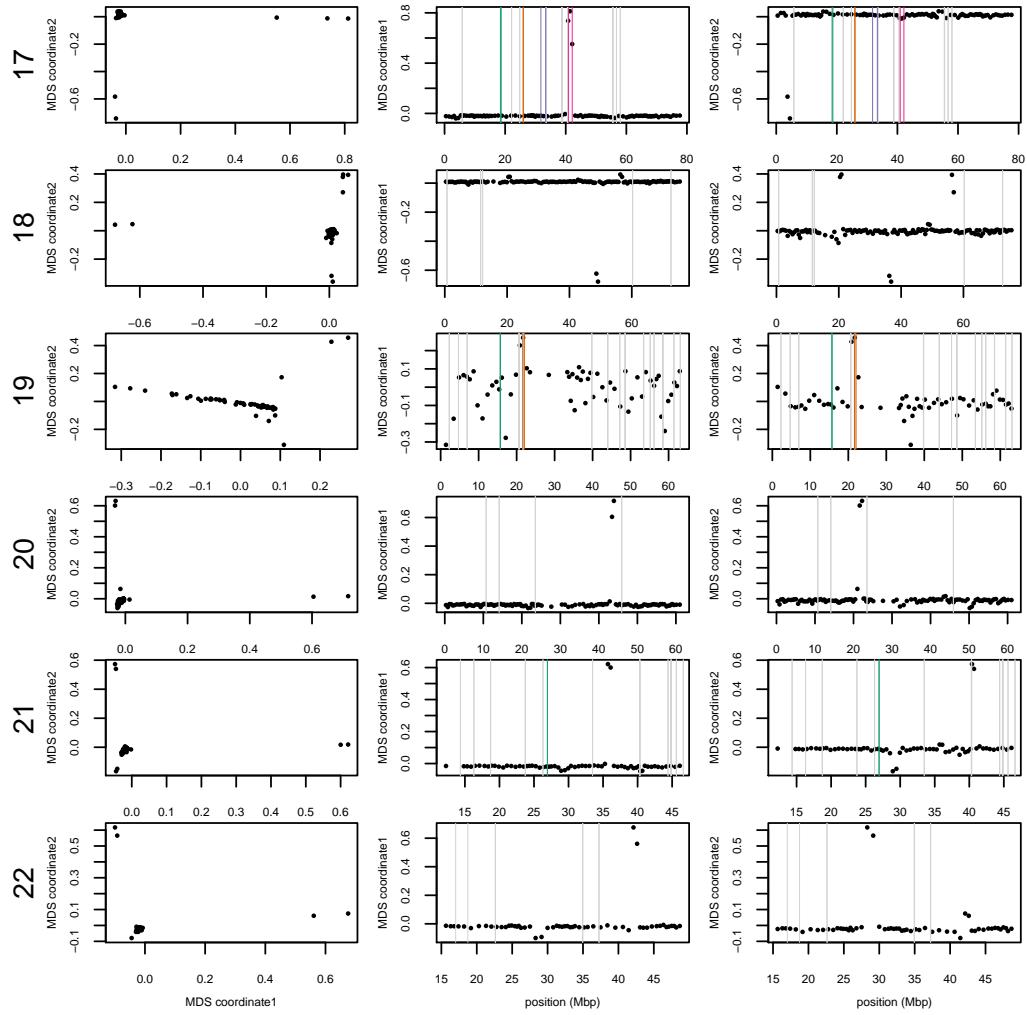


Figure S9: MDS plots for human chromosomes 17-22, as in Supplemental Figure S7.

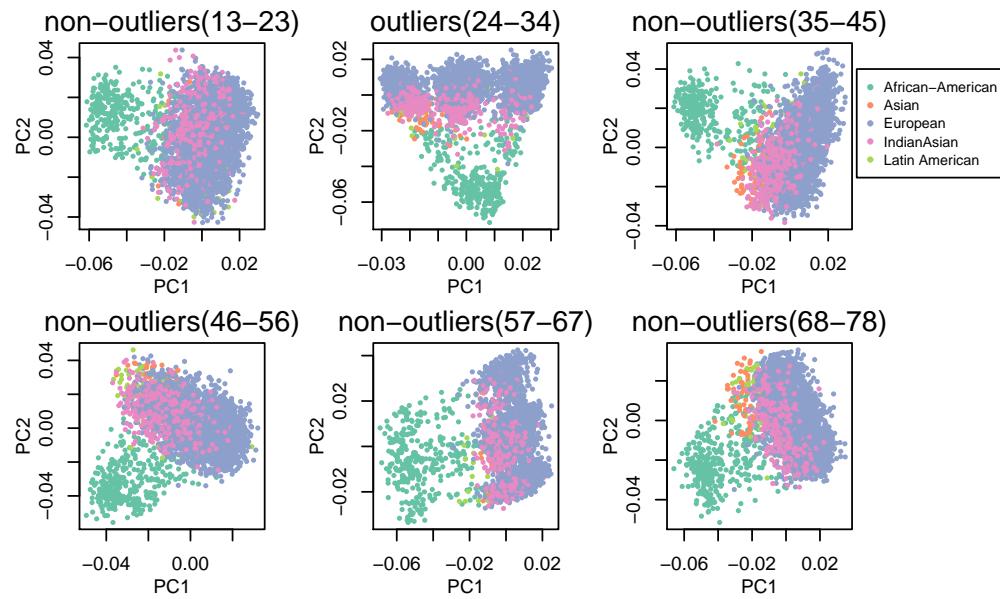


Figure S10: Comparison of PCA figures within outlying windows (center column) and flanking non-outlying windows (left and right columns) for the two windows having outlying MDS scores on chromosome 8.

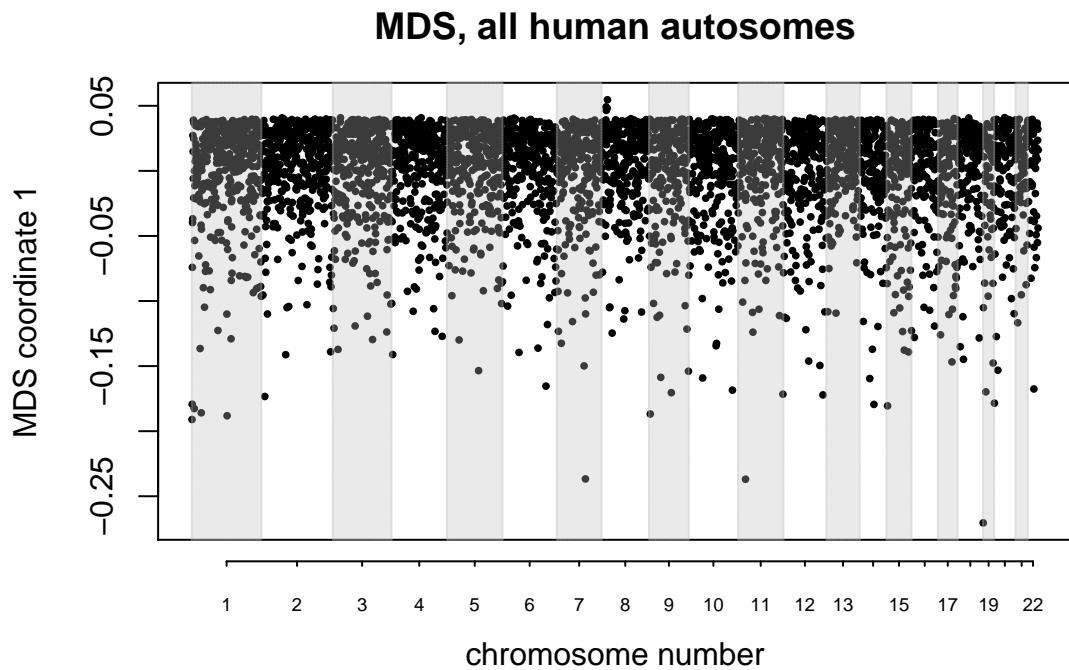


Figure S11: MDS visualization of variation in the effects of population structure amongst windows across *all* human autosomes simultaneously. The small group of windows with positive MDS values lie around the inversion at 8p23.

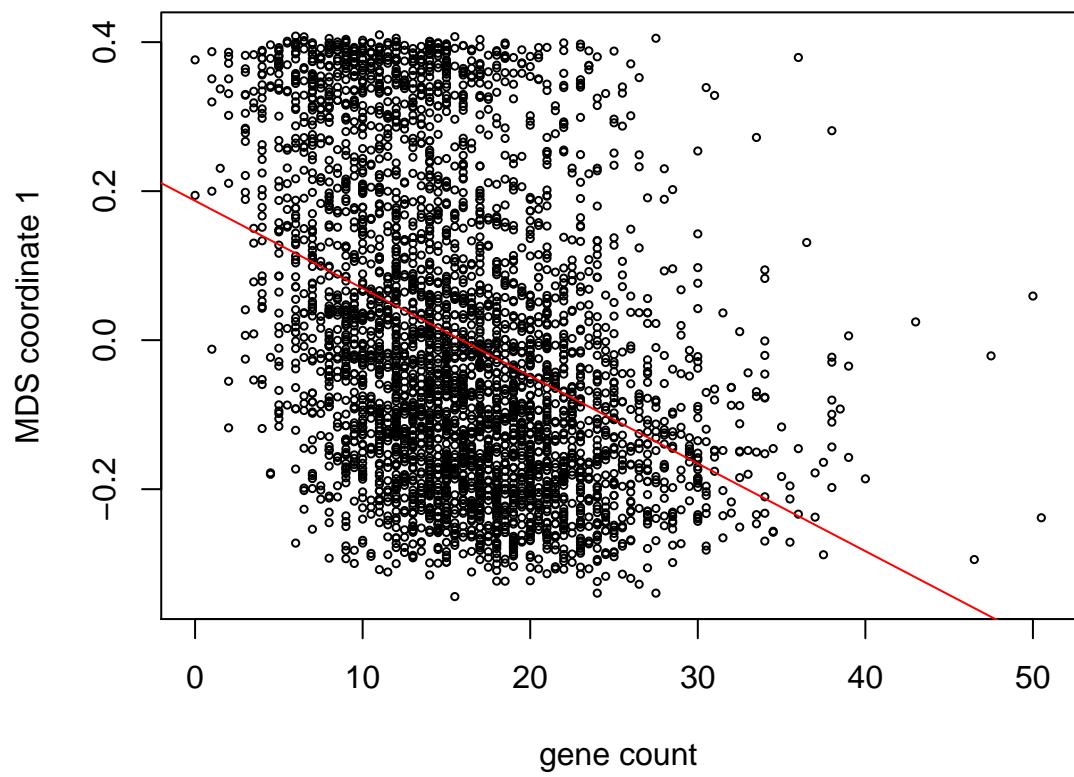


Figure S12: First MDS coordinate against gene density for all 8 chromosomes of *M. truncatula*. The first MDS coordinate is significantly correlated with gene count ($r = 0.149$, $p = 2.2 \times 10^{-16}$).

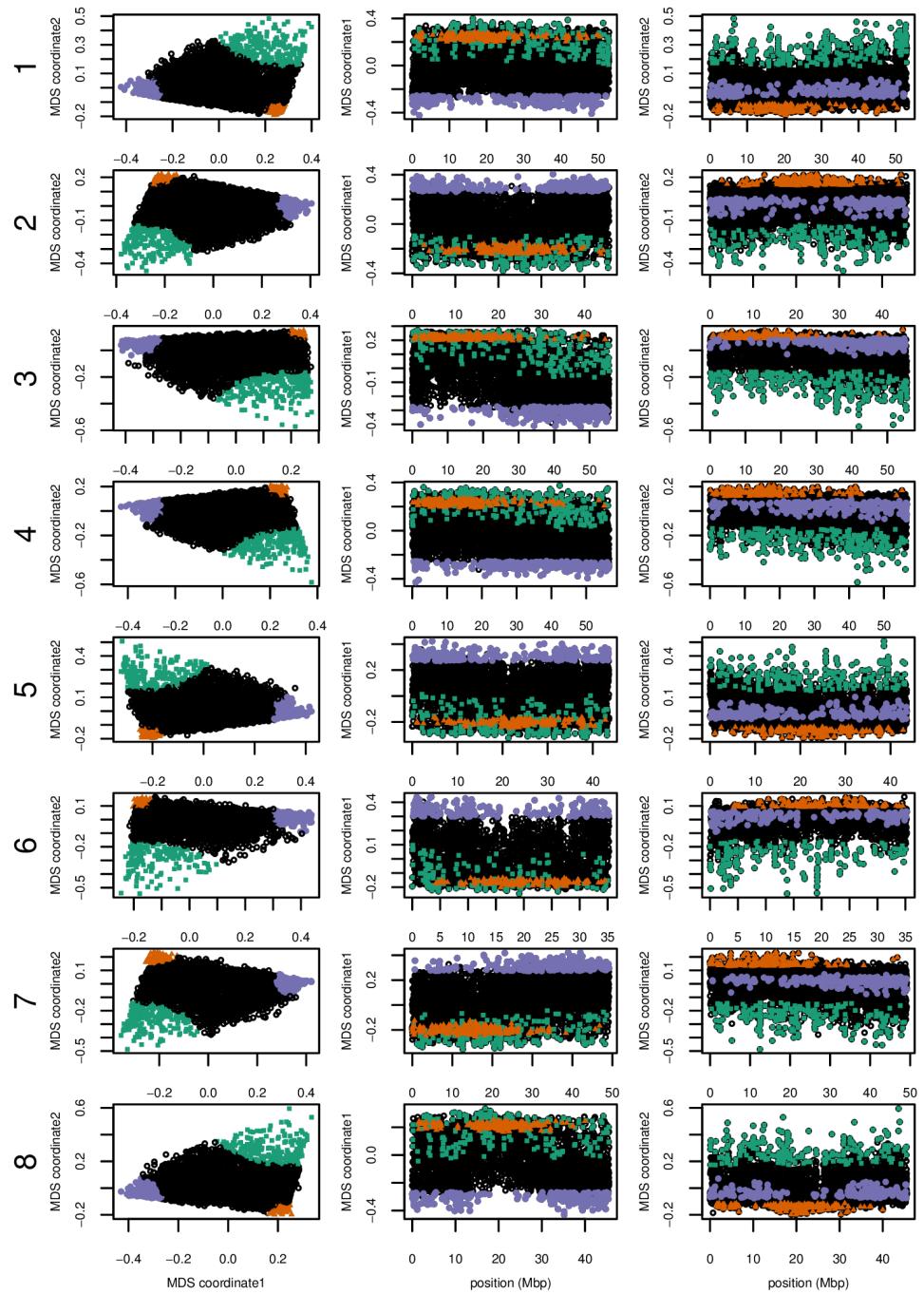


Figure S13: MDS visualizations of the effects of population structure for all 8 chromosomes of the *Medicago truncatula* data.

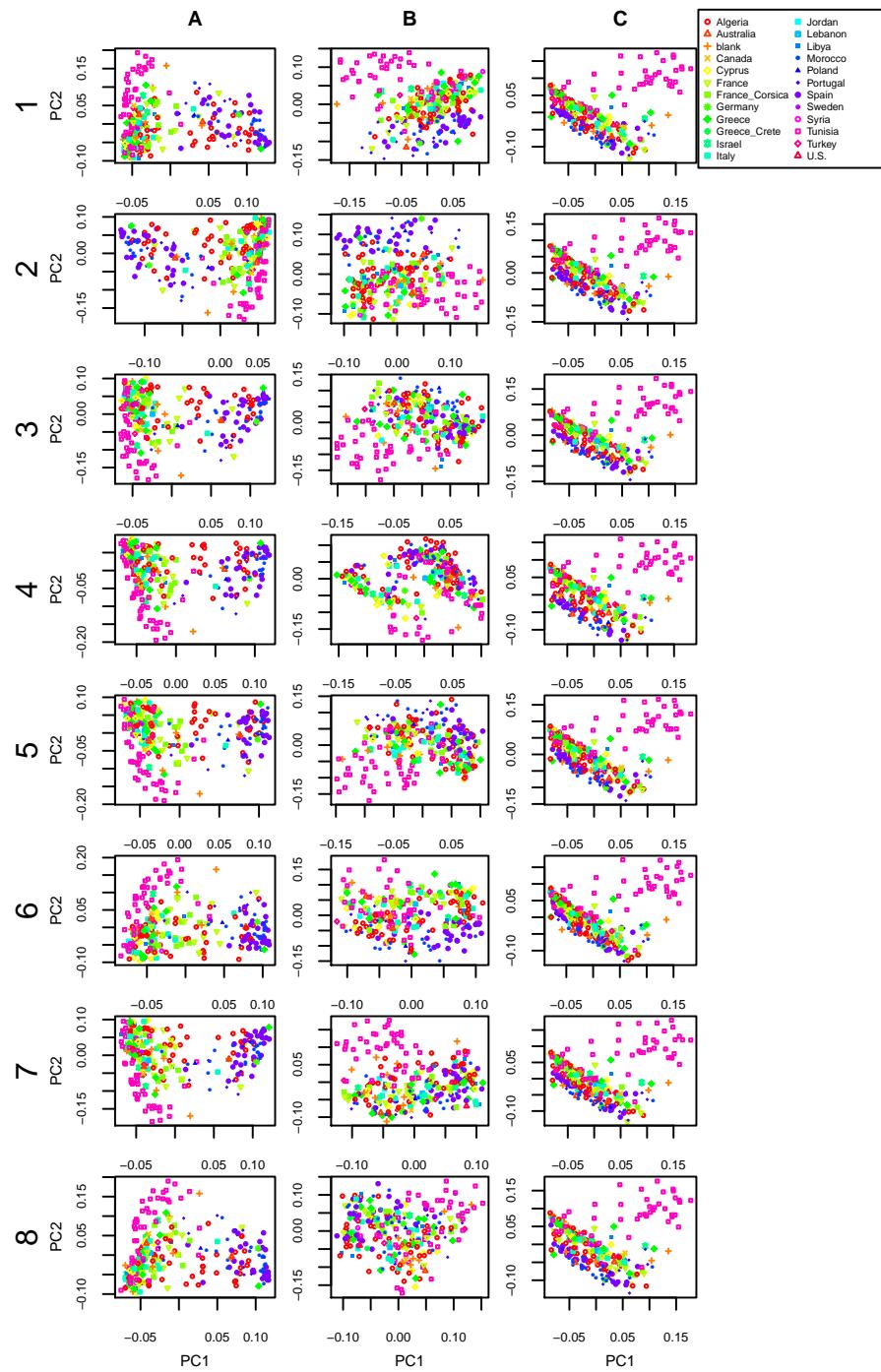


Figure S14: PCA plots for regions colored in Figure S13 on all 8 chromosomes of *Medicago truncatula*: (A) green, (B) orange, and (C) purple.

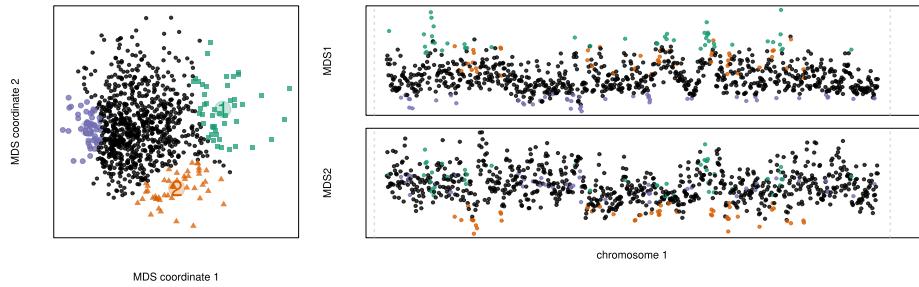


Figure S15: MDS visualizations of simulation run 012720 (RECOMBTYPE, “step”; SELTYPE, “neutral”)

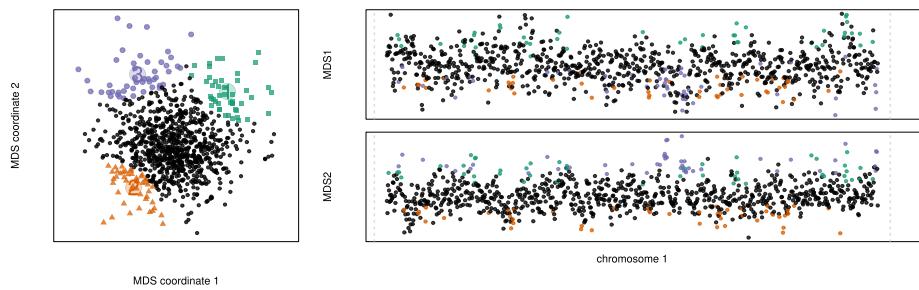


Figure S16: MDS visualizations of simulation run 027034 (RECOMBTYPE, “base”; SELTYPE, “neutral”)

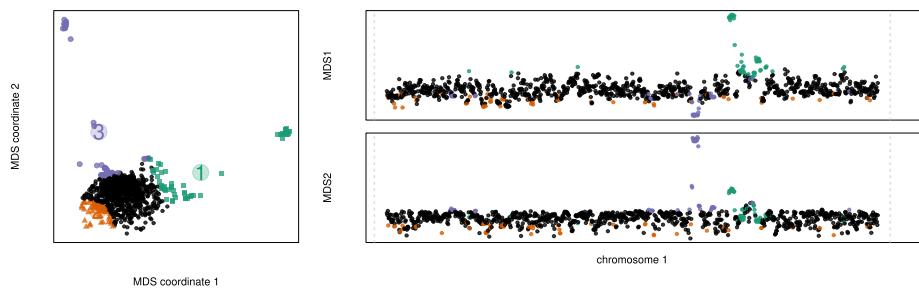


Figure S17: MDS visualizations of simulation run 015598 (RECOMBTYPE, “hotspot”; SELTYPE, “neutral”)

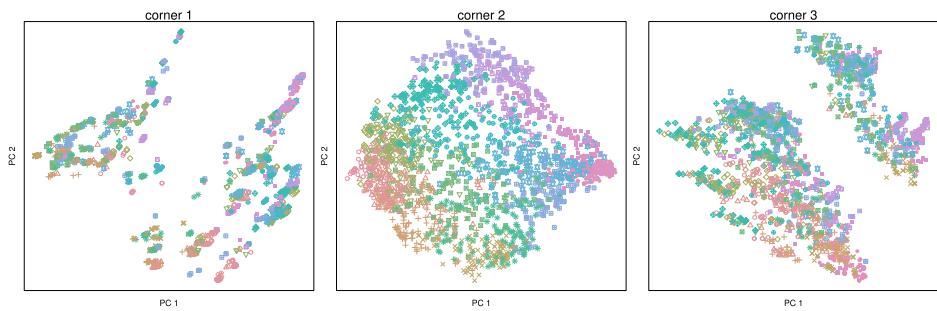


Figure S18: PCA plots for regions colored in Figure S17

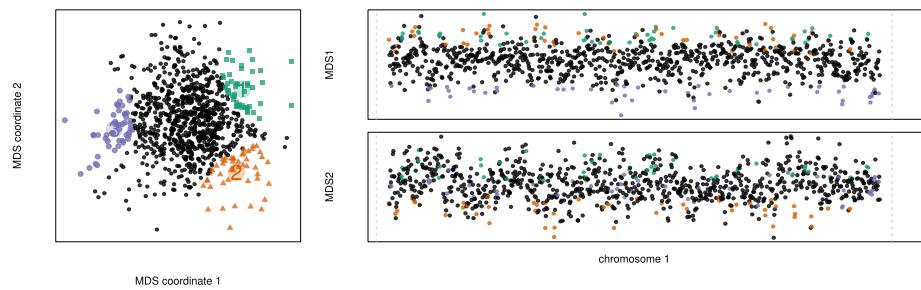


Figure S19: MDS visualizations of simulation run 005464 (RECOMBTYPE, “base”; SELTYPE, “selected”)

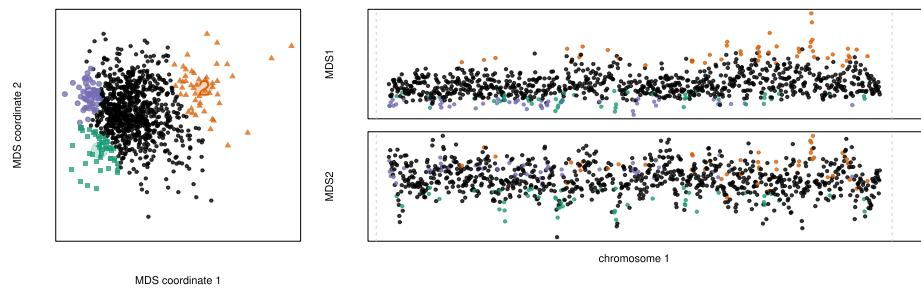


Figure S20: MDS visualizations of simulation run 031486 (RECOMBTYPE, “base”; SELTYPE, “selected”)

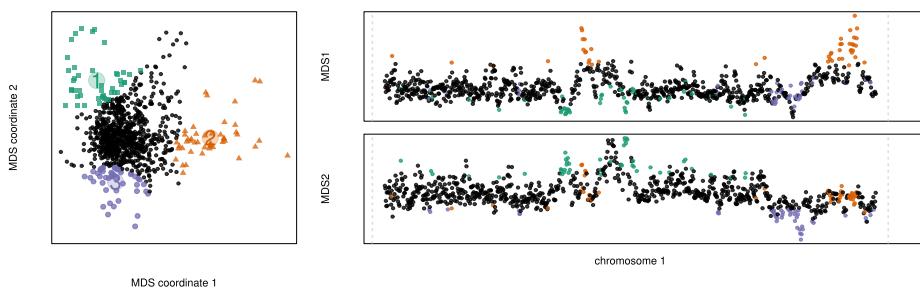


Figure S21: MDS visualizations of simulation run 009673 (RECOMBTYPE, “step”; SELTYPE, “selected”)

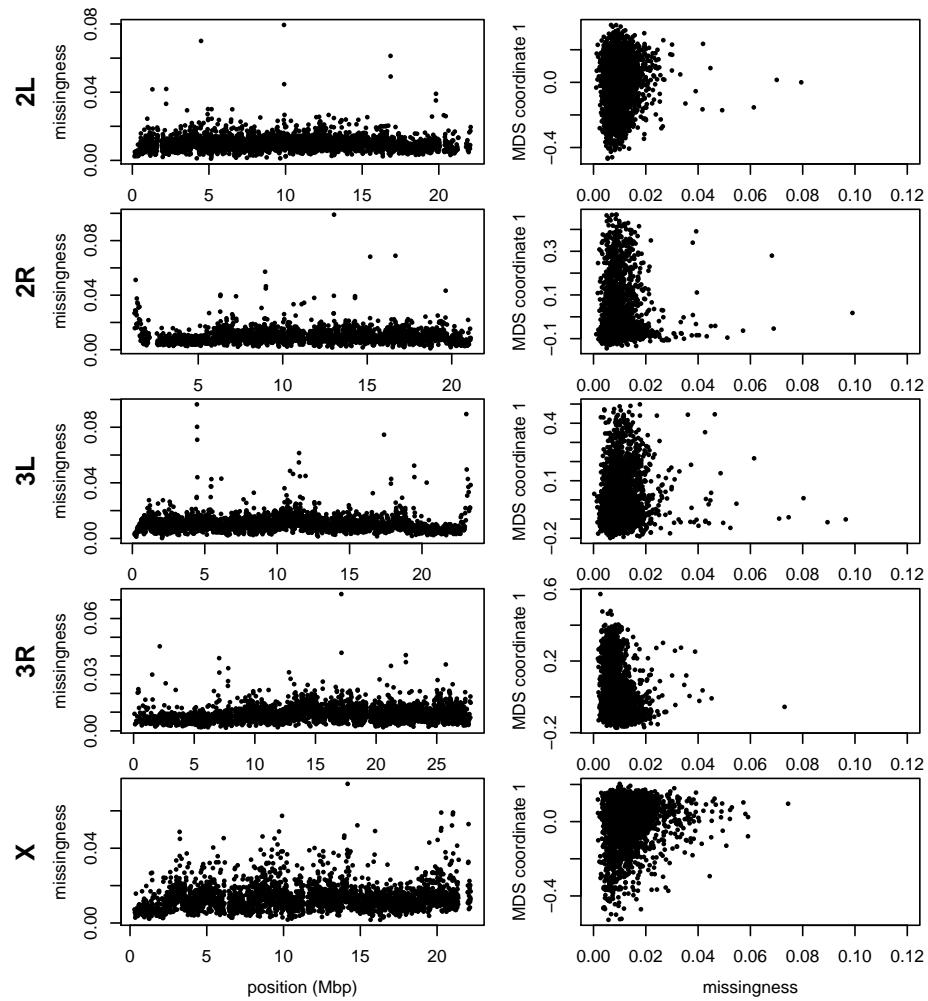


Figure S22: The missingness ratio along genome and the correlation of MDS against missingness in Drosophila.

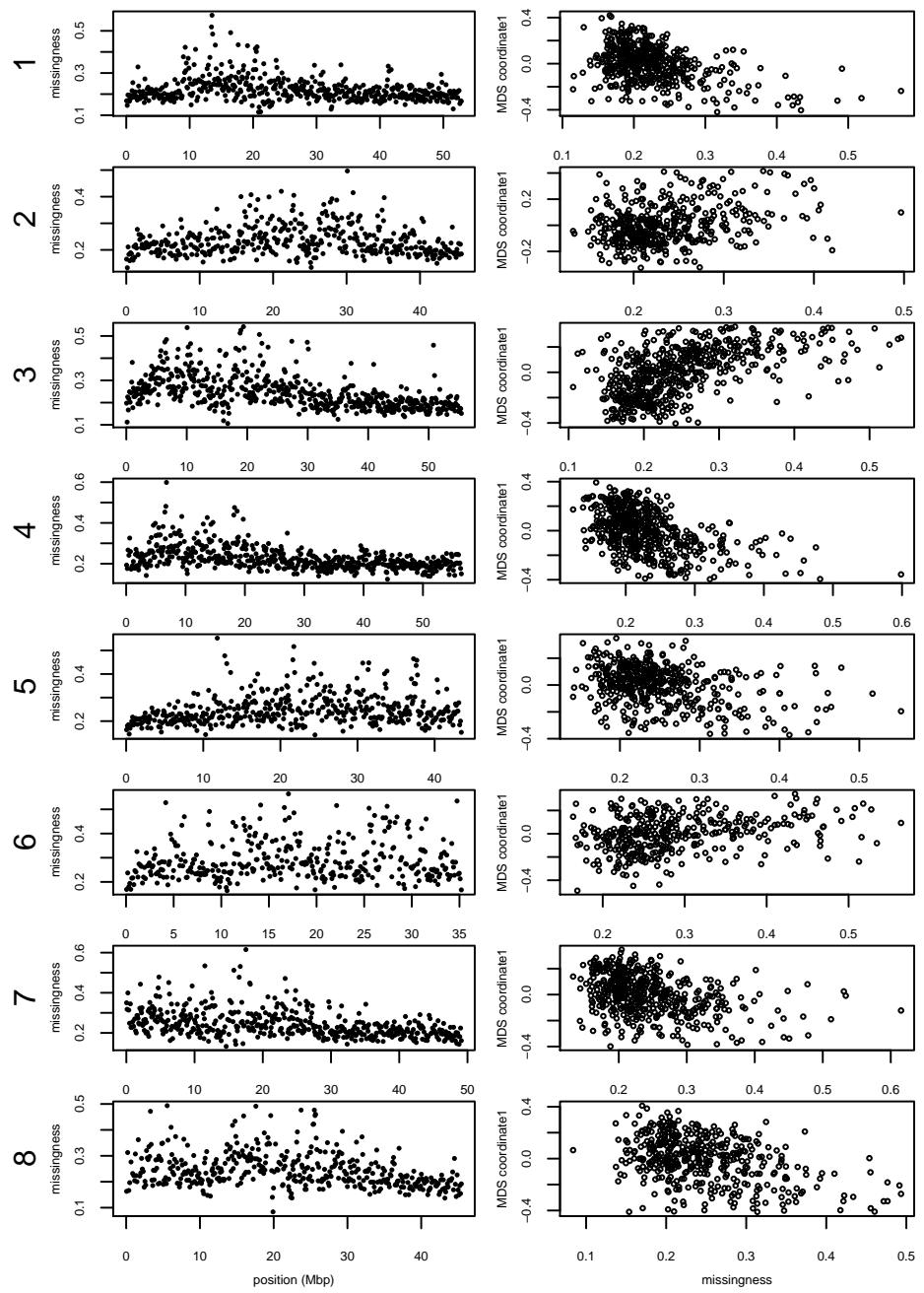


Figure S23: The missingness ratio along genome and the correlation of MDS against missingness in *Medicago*.

Resubmission Cover Letter
Genetics

Han Li
and Peter Ralph
August 9, 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.

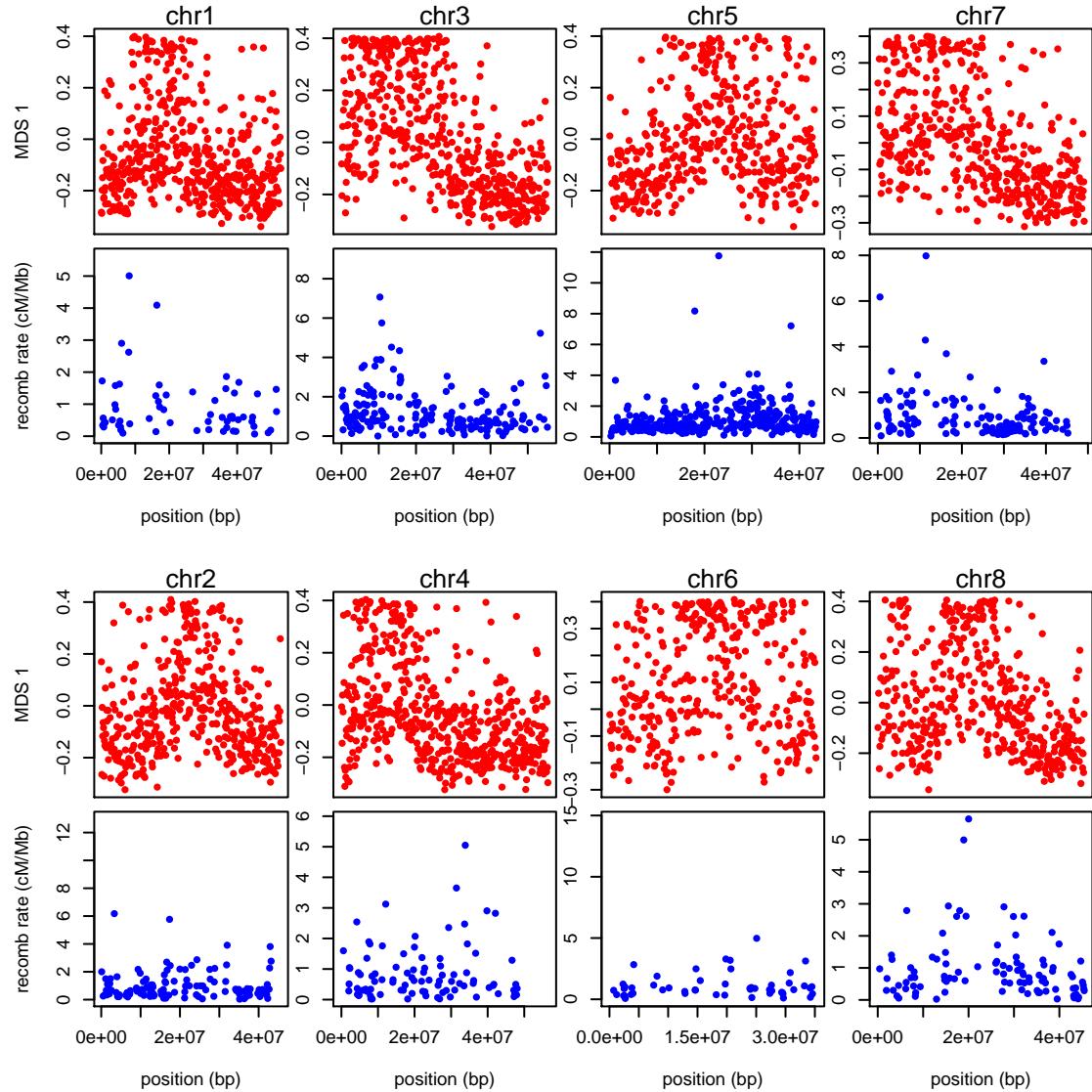
Reply: This is a good point, but XXX We now cite PCAdmix (p. ??, l. ??).

(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.