Local PCA Shows How Population Structure Differs Along the Genome

Han Li, Peter Ralph

August 1, 2016

Abstract

Dimension reduction techniques such as principal component analysis (PCA) are often used to discover and visualize the patterns of mean kinship between individuals in large genomic datasets, known as population structure. Genome-wide mean kinship is an average of the relationships across all locus-specific genealogical trees, which can be strongly affected on intermediate genomic scales by many biological factors, such as linked selection. We show how to use local principal components analysis to describe this meso-scale variation, and apply the method to genomic data from three species, finding in each that population structure varies on the scale of megabases to tens of megabases. In a global human dataset, small, discontinuous variation is likely explained by polymorphic chromosomal inversions. In a dataset of primarily African *Drosophila melanogaster*, large, continuous variation across each chromosome arm is explained by known chromosomal inversions thought to be under recent selection. In a range-wide dataset of *Medicago truncatula*, common axes of variation in population structure are shared between chromosomes, correlate with local gene density, and may be caused by background selection or local adaptation. The method is implemented as an R package, called lostruct.

Add percent of variance explained by first two MDS coordinates, after subtracting the genome-wide mean covariance matrix.

Add references to supplemental whole-genome PCA plots in the text.

1 Introduction

The phrase "population structure", used to describe patterns in genome-wide mean kinship, calls to mind to reduced gene flow between subpopulations. However, it is widely recognized that because of selection the effects of gene flow are not equal everywhere on the genome, and patterns of polymorphism and divergence can vary significantly depending on factors including local gene density. This implies that, counterintuitively, the population structure of a species depends on which part of the genome is being examined.

Realized kinship summarizes the shapes of the genealogical trees at each location along the genome. Since these trees vary along the genome, so does realized kinship; but averaging over sufficiently many trees we hope to get a stable estimate, independent of the genomic region chosen. This hope is not entirely justified: for instance, kinship on sex chromosomes is expected to differ from the autosomes; and positive or negative selection on particular loci can dramatically disort shapes of nearby genealogies (Barton 2000; Charlesworth 2012; Maynard Smith and Haigh 1974; Neher 2013). Indeed, many species show chromosome-scale variation in diversity and divergence (sometimes dramatic: Langley et al. 2012); species phylogenies can differ along the genome due to incomplete lineage sorting (e.g., Pease and Hahn 2013), adaptive introgression and/or local adaptation (e.g., Ellegren et al. 2012; Nadeau et al. 2012; Pool 2015; Vernot and Akey 2014); and theoretical expectations predict that geographic patterns of relatedness should depend on selection (Charlesworth et al. 2003). Nonetheless, it is not generally known to what extent patterns of kinship vary along the genome, nor for what reasons.

Patterns in genome-wide kinship are often summarized by applying principal components analysis (PCA, Patterson et al. 2006) to the genetic covariance matrix, as pioneered by Menozzi et al. (1978). The results of PCA can be related to the genealogical history of the samples, such as time to most recent common ancestor and migration rate between populations (McVean 2009; Novembre and Stephens 2008), and sometimes produce "maps" of population structure that reflect the samples' geographic origin distorted by rates of gene flow (Novembre et al. 2008).

Furthermore, modeling "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 is essential for doing GWAS in species with substantial variation in population structure,

In this work we describe variation in population structure using local PCA, where "local" refers to proximity along the genome. A number of general methods for dimensionality reduction also use a strategy of "local PCA" (e.g., Kambhatla and Leen 1997; Manjón et al. 2013; Roweis and Saul 2000; Weingessel and Hornik 2000), performing PCA not on the entire dataset but instead on subsets of observations, providing local pictures which are then stitched back together to give a global picture. At first sight, this differs from our method in that we restrict to subsets of *variables* instead of subsets of observations. However, if we flip perspectives and think of each genetic variant as an observation, our method shares common threads, although the ultimate goals and methods for visualization are different. Future methods for visualization of genomic data may benefit from other advances in this substantial literature (reviewed in Van Der Maaten et al. 2009).

2 Methods

The general steps in our method could be carried out in many ways. As depicted in Figure 1, these are (1) divide the genome into windows, (2) summarize the patterns of relatedness (population structure) in each window, (3) measure dissimilarity in population structure between each pair of windows, (4) visualize the resulting dissimilarity matrix, and (5) combine similar windows to more accurately estimate local population structure. Details of how we carried these out are given below.

2.1 Datasets

We used three publically available datasets (summarized in Table 1); we converted each to a numeric matrix (with one row per polymorphic variant and one column per sample) by replacing each genotype with the number of nonreference alleles (or NA for missing data). A normalization step (see below) ensures the result does not depend on the choice of reference allele.

***Drosophila melanogaster*:** We used whole-genome sequencing data from the Drosophila Genome Nexus (`http://www.johnpool.net/genomes.html`, Lack et al. 2015), consisting of the Drosophila Population Genomics Project phases 1–3 (Langley et al. 2012; Pool et al. 2012), and additional African genomes (Lack et al. 2015). After removing 20 genomes with more than 8% missing data, we were left with 380 samples from 16 countries across Africa and Europe. Since the *Drosophila* samples are from inbred lines or haploid embryos, we treat the samples as haploid when recoding; regions with residual heterozygosity were marked as missing in the original dataset; we also removed positions with more than 20% missing data. Each chromosome arm we investigated (X, 2L, 2R, 3L, and 3R) has 2–3 million SNPs; PCA plots for each arm are shown in Figure S8.

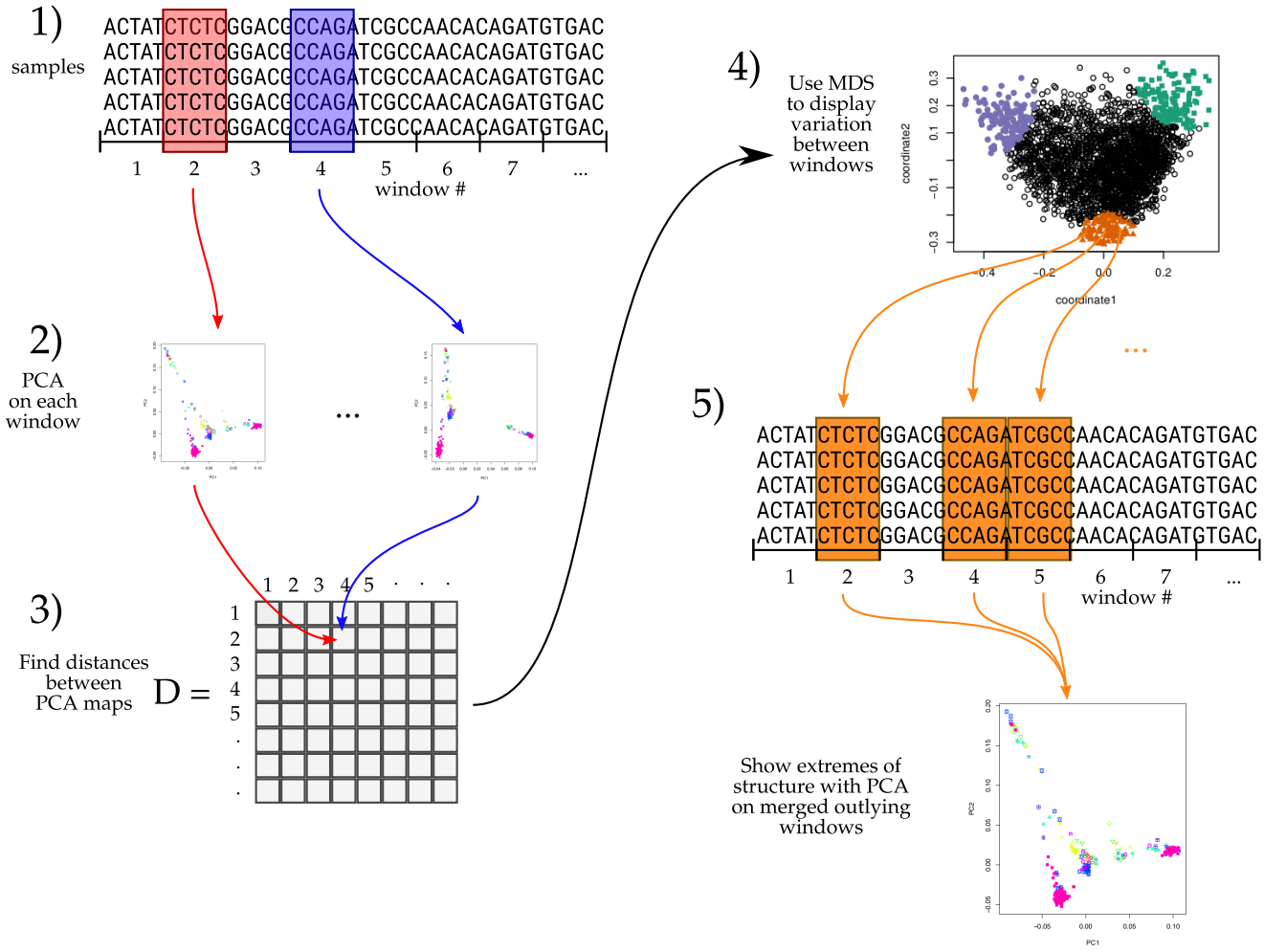

Figure 1: An illustration of the method; see text for details.

Human: We also used genomic data from the entire POPRES dataset (Nelson et al. 2008), which has array-derived genotype information for 447,267 SNPs across the 22 autosomes of 3,965 samples in total: 346 African-Americas, 73 Asians, 3,187 Europeans and 359 Indian Asians. Since these data derive from genotyping arrays, the SNP density is much lower than the other datasets, which are each derived from whole genome sequencing. We excluded the sex chromosomes and the mitochondria. PCA plots for each chromosome, separately, are shown in Figure S9.

***Medicago truncatula*:** Finally, we used whole-genome sequencing data from the *Medicago truncatula* Hapmap Project (Tang et al. 2014), which has 263 samples from 24 countries, primarily distributed around the Mediterranean basin. Each of the 8 chromosomes has 3–5 million SNPs; PCA plots for these are shown in Figure S10. we did not use the mitochondria or chloroplasts.

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

Table 1: Descriptive statistics for each dataset used. Columns 2–4 give statistics describing the window sizes used for most analyses for each dataset (windows had a fixed number of SNPs). The final column gives the percent variance explained by the top two principal components, averaged across independent PCA of each window.

2.2 PCA in genomic windows

After recoding, we divided the genome into contiguous segments ("windows"; see below), and applied Principal Component Analysis (PCA) as described in McVean (2009) separately to the submatrices that corresponded to each window. Precisely, denote by Z the

$L \times N$ recoded genotype matrix (L is the number of SNPs and N is the sample size), and by $\overline{Z_s}$ the mean of non-missing entries for allele s, so that $\overline{Z_s} = \frac{1}{n_s} \sum_j Z_{sj}$, where the sum is over the n_s nonmissing genotypes. We first compute the mean-centered matrix X, as $X_{si} = Z_{si} - \overline{Z_s}$, and preserving missingness. (This mean-centering makes the result not depend on the choice of reference allele, exactly if there is no missing data, and approximately otherwise.) Next, we find the covariance matrix of X, denoted C, as $C_{ij} = \frac{1}{m_{ij}-1} \sum_s X_{si} X_{sj} - \frac{1}{m_{ij}(m_{ij}-1)} (\sum_s X_{si})(\sum_s X_{sj})$, where all sums are over the m_{ij} sites where both sample i and sample j have nonmissing genotypes. The principal components are the eigenvectors of C, normalized to have Euclidean length equal to one, and ordered by magnitude of the eigenvalues.

The top few principal components generally display population structure; we usually use the first two (referred to as $PC1$ and $PC2$). The above procedure can be performed on any subset of the data; for future reference, denote by $PC1_j$ and $PC2_j$ the result after applying to all SNPs in the j^{th} window.

Since eigenvectors are still only defined up to sign, when comparing between windows we choose signs to best match each other: after choosing $PC1_1$, for instance, if u is the first eigenvector obtained from the covariance matrix for window j, then we next choose $PC1_j = \pm u$, where the sign is chosen according to which of $\|PC1_1 - u\|$ or $\|PC1_1 + u\|$ is smaller.

Several of the datasets we use have unbalanced representations of diverged populations, which can have a strong impact on the results of PCA. (The principal axes may describe variation *within* an overrepresented group rather than more significant variation between groups.) Therefore, to check the effect of sampling on our results, we used a variant of PCA that gives roughly equal weight to each group of samples, rather than to each sample. The rationale and implementation of this method are described in Appendix A.

2.3 Choosing window length

The choice of window length entails a balance between signal and noise. In very short windows, genealogies of the samples will only be represented by a few trees, so variation between windows represents demographic noise rather than meaningful variation in population structure. Longer windows generally have more distinct trees (and SNPs), allowing for less noisy estimation of local population structure. However, to better resolve meaningful signal, i.e., differences in population structure along the genome, we would like reasonably short windows.

Since we summarize population structure using relative positions in the principal component maps, we quantify "noise" as the standard error of a sample's position on PC1 in a particular window, averaged across windows and samples, and "signal" as the standard deviation of the sample's position on PC1 over all windows, averaged over samples. (Recall that the signs for PCs are chosen to match each other.) Then, the mean variance across windows is

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

where $PC1_{ij}$ is the position of the i^{th} individual on $PC1$ in window j, and $\overline{PC1}_j = (1/N)\sum_{j=1}^{N} PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jackknife (Busing et al. 1999; Efron 1982): we divide the j^{th} window into 10 equal-sized pieces, and let $PC1_{ij,k}$ denote the first principal component of this region found after removing the k^{th} piece; then the estimate of the squared standard error is $\sigma^2_{ij} =$

$\frac{9}{10}\sum_{k=1}^{10}(PC1_{ij,k} - \frac{1}{10}\sum_{\ell=1}^{10} PC1_{ij,\ell})^2$. Averaging over samples and windows,

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

For the main analysis, we defined windows to each consist of the same number of neighboring SNPs, and calculated σ^2_{signal} and σ^2_{noise} for a range of window sizes (i.e., numbers of SNPs). For our main results we chose the smallest window for which σ^2_{signal} was consistently larger than σ^2_{noise} (but checked other sizes); the values for various window sizes across *Drosophila* chromosomes are shown in Table 2. In the cases we examined, we found nearly identical results after varying window size, and choosing windows to be of the same 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^2_{noise}	2.05	1.64	1.18	0.17	0.04
	σ^2_{signal}	2.76	2.69	2.23	0.68	0.31
2R	σ^2_{noise}	2.18	1.92	1.63	0.58	0.13
	σ^2_{signal}	2.78	2.70	2.65	2.31	1.82
3L	σ^2_{noise}	2.08	2.00	1.64	0.73	0.25
	σ^2_{signal}	2.60	2.52	2.40	1.68	1.89
3R	σ^2_{noise}	1.95	1.76	1.44	0.59	0.20
	σ^2_{signal}	2.58	2.51	2.44	1.96	1.40
X	σ^2_{noise}	2.48	2.04	1.54	1.62	0.17
	σ^2_{signal}	2.61	2.43	2.30	0.32	1.14

Table 2: 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).

2.4 Similarity of population structure between windows

We think of local population structure as being summarized by *relative* position of the samples in the space defined by the top principal components. Since rotations of this space imply identical population structure, we do not compare population structure of different genomic windows using the PCs directly, but instead compare the low-dimensional approximations of the local covariance matrices obtained using the top k PCs. (For results shown here, we use $k = 2$; results using larger numbers of PCs were nearly identical.) We do this, rather than using the entire covariance matrix, both for computational efficiency and because the top PCs summarize important population structure, so using only these should reduce the effect of noise. To remove the effect of variation in mutation rate between windows, we also rescale each covariance matrix so that the underlying data matrix has trace norm equal to one.

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 equal to the ℓ^{th} princpal component of the i^{th} window, multiplied by $\sqrt{\lambda_{\ell i}/\sum_{m=1}^{k}\lambda_{mi}}$, where $\lambda_{\ell i}$ is the ℓ^{th} eigenvalue of the genetic covariance matrix. Then, the rescaled, rank k approximate covariance matrix for the i^{th} window is

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

We then use Euclidean distance D_{ij} between the matrices $M(i)$ and $M(j)$ to measure the similarity of population structure for the i^{th} window and j^{th} window. Since $\sum_{ij}(A_{ij} - B_{ij})^2 = \sum_{ij}(A_{ij}^2 + B_{ij}^2) - 2\,\mathrm{tr}(A^T B)$, and $M(i) = V(i)V(i)^T$, then due to the orthogonality of eigenvectors and the cyclic invariance of trace, D_{ij} can be computed efficiently as

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

2.5 Visualization of results

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

We then locate variation in population structure along the genome by choosing collections of windows that are nearby in MDS coordinates, and map their positions along the genome. Their common population structure is visualized by extracting the corresponding genomic regions and performing PCA on all, aggregated, regions.

2.6 Software

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

3 Results

In these three species, PCA plots vary along the genome in a systematic way, showing strong chromosome-scale correlations. This implies that variation is due to meaningful variation in population structure, since fluctuations due to demographic noise are not expected to show long distance correlations. Below, we discuss the results and likely underlying causes.

3.1 *Drosophila melanogaster*

We ran the method on chromosome arms 2L, 2R, 3L, 3R and X separately. For each, the two-dimensional MDS visualization resembles a triangle, as seen in Figure 2. Since the relative position of each window in this plot shows the similarity between windows, this suggests that there are at least three extreme types of population structure typified by windows found in the three corners of the triangle, and that other windows' population structure may be a mixture of those extremes.

Corbett-Detig and Hartl (2012) says In(1)Be is very young, ∼ 60 years, probably why it doesn't show up. Also cite Corbett-Detig et al. (2012) for a computational method for detecting inversions.

To investigate these extremes, for each chromosome arm we visually selected three "extreme" windows in the MDS plot, and take out the 5% of windows that are closest to it in the MDS coordinates; these three groups of windows are shown in color in Figure 2 and PCA plots for each collection are shown in Figure 3.

The variation in population structure turns out to be almost entirely explained by several large inversions that are polymorphic in these samples, disussed in Corbett-Detig and Hartl (2012) and Langley et al. (2012). To depict this, Figure 4 shows the PCA plots in Figure 3 recolored by the orientation of the inversion for each sample. Taking chromosome arm 2L as an example, the two regions of similar, extreme population structure shown in green in the first row of Figure 2 lie directly around the breakpoints of the inversion In(2L)t, and the PCA plots in the first rows of Figure 4 shows that population structure here is mostly determined by inversion orientation. The regions shown in purple on chromosome 2L lie near the centromere, and have population structure reflective of two axes of variation, seen in Figure 3, which correspond roughly to latitude within Africa and to degree of cosmopolitan admixture respectively (see Lack et al. (2015) for more about admixture

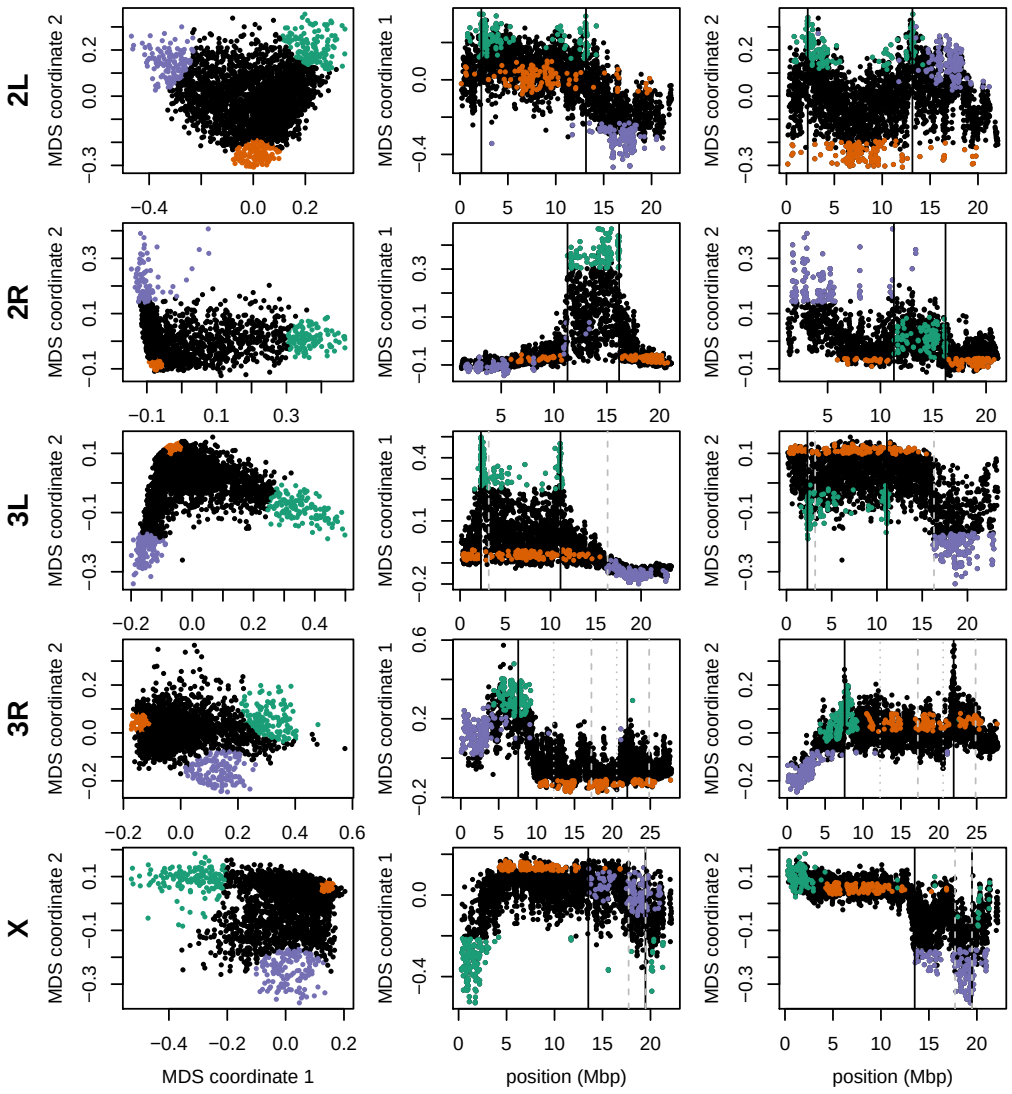

Figure 2: Variation in structure for windows across *Drosophila melanogaster* chromosome arms. In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window 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. Solid black lines are for the inversions we used in Figure 4, while dotted grey lines are for other known inversions.

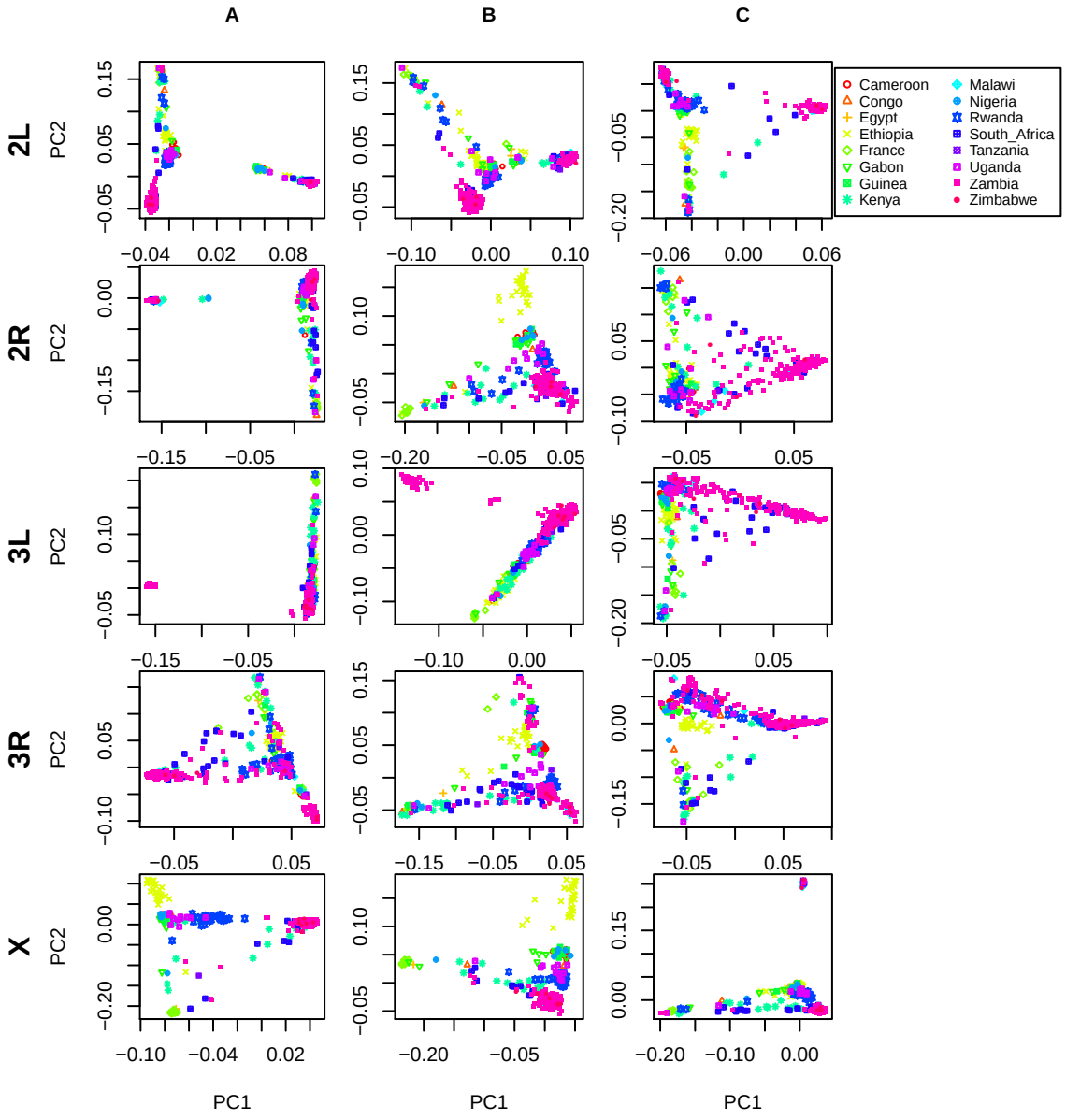

Figure 3: PCA plots for the three sets of genomic windows colored in Figure 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Figure 2; the second is for orange windows; and third is for purple windows.

in this sample). The regions shown in orange on chromosome 2L mostly lie inside the inversion, and show population structure that is a mixture between the other two, as expected due to recombination within the (long) inversion (Guerrero et al. 2011). Similar results are found in other chromosome arms, albeit complicated by the coexistence of more than one polymorphic inversion; however, each breakpoint visibly affects patterns in the MDS coordinates.

3.2 Human

As for the *Drosophila* data, we ran our method separately on all 22 human autosomes. On each, variation in population structure was dominated by a small number of windows having similar population structure to each other that differed dramatically from the rest of the chromosome. These may be primarily inversions: outlying windows coincide with three of the six large polymorphic inversions described in Antonacci et al. (2009), notably a particularly large, polymorphic inversion on 8p23 (Figure 5). Similar plots for all chromosomes are shown in Supplementary Figures S2, S3, S4. PCA plots of many outlying windows show a characteristic trimodal shape (shown for chromosome 8 in Figure S5), presumably distinguishing samples having each of the three diploid genotypes for each inversion orientation (although we do not have data on orientation status). This trimodal shape has been proposed as a method to identify inversions (Ma and Amos 2012), but distinguishing this hypothesis from others, such as regions of low recombination rate, would require additional data.

We also ran the method grouping all 22 autosomes together, and found that, remarkably, the inversion on chromosome 8 is still the most striking outlying signal (Figure S1). Further investigation with a denser set of SNPs, allowing a finer genomic resolution, may yield other patterns.

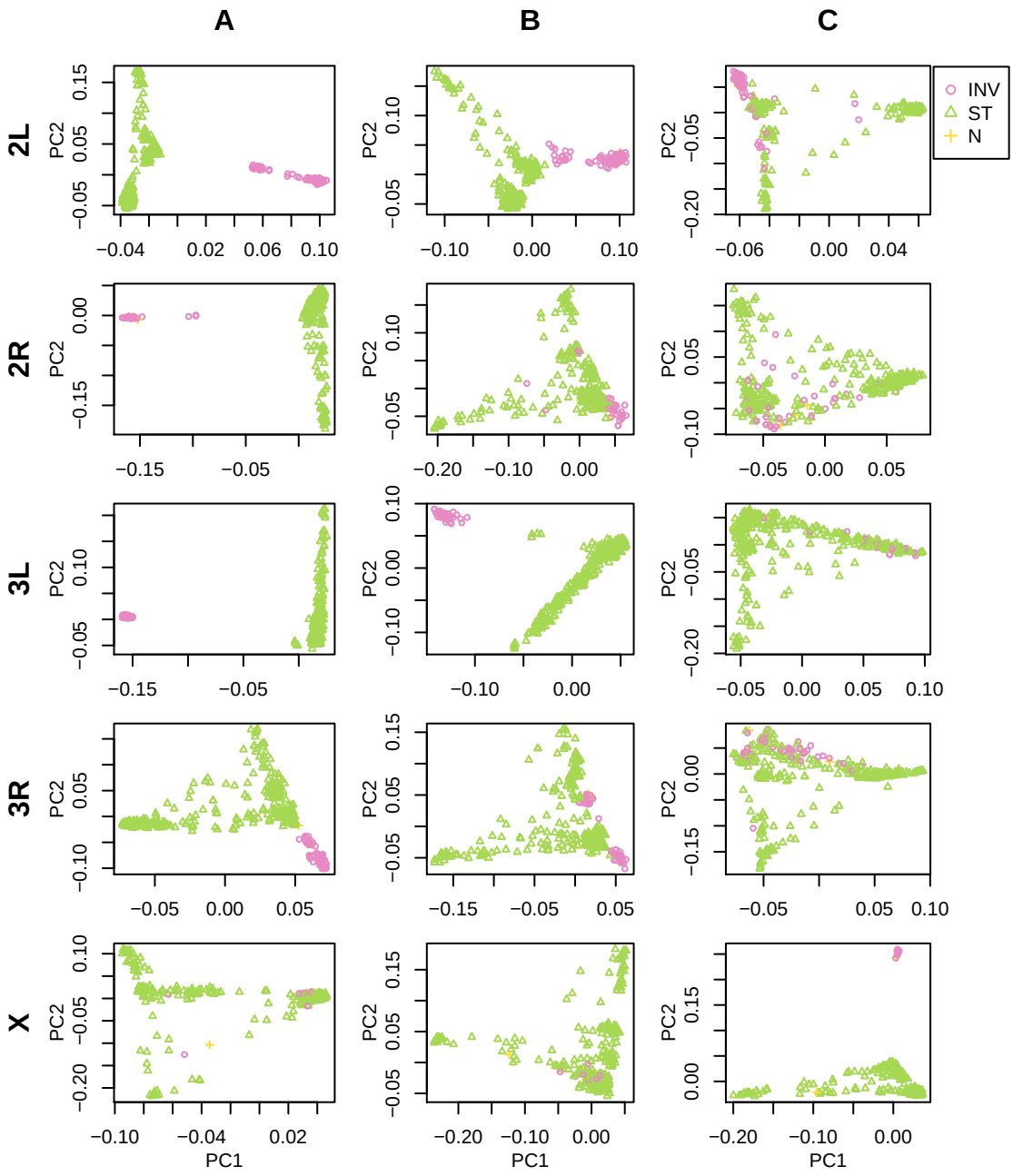

Figure 4: As in Figure 2, except that samples are colored by orientation of the corresponding polymorphic inversions, In(2L)t, In(2R)NS, In(3L)OK, In(3R)K and In(1)A respectively. (data from Lack et al. 2015)

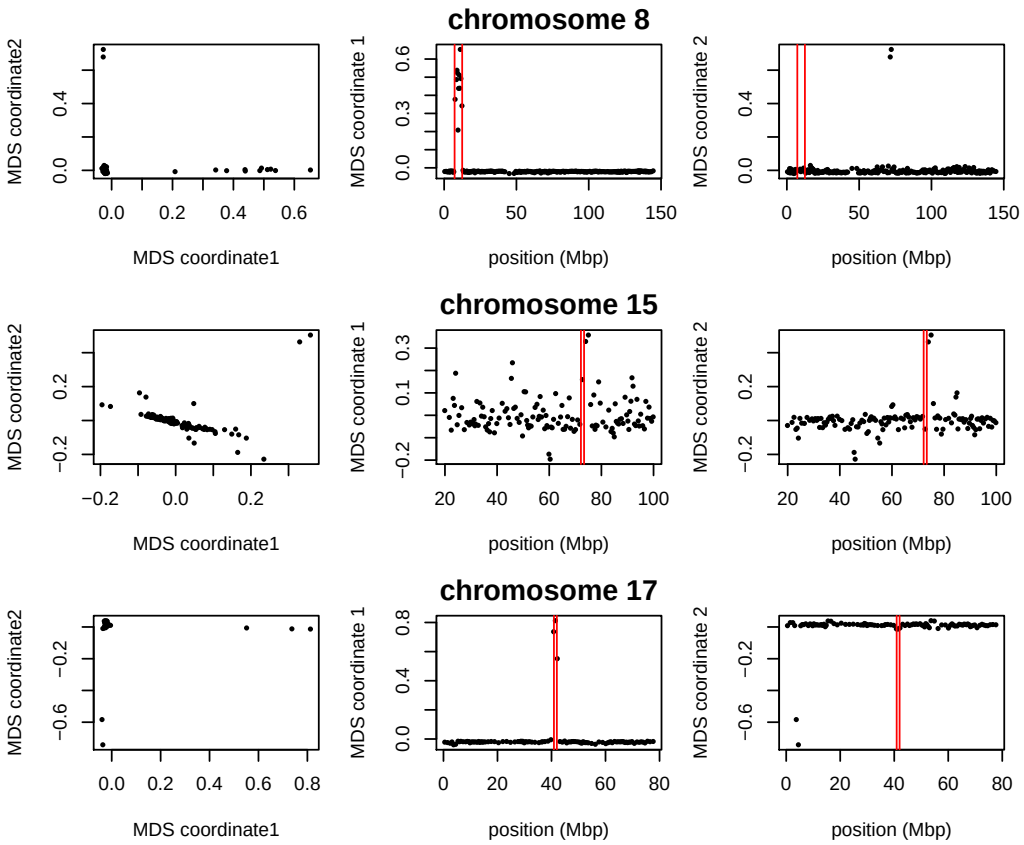

Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in each plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position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

3.3 *Medicago truncatula*

Unlike the other two species, the method applied separately on all eight chromosomes of *Medicago truncatula* showed similar patterns of gradual change in population structure across each chromosome, with no indications of chromosome-specific patterns. This consistency suggests that the factor driving the population structure for each chromosome is the same, as might be caused by varying strengths of linked selection. To verify that variation in population structure is shared across chromosomes, we applied the method to all chromosomes together. Results for chromosome 3 are shown in Figures 6 and 7, and other chromosomes are similar: across chromosomes, the high values of the first MDS coordinate coincide with the position of the heterochromatic regions surrounding the centromere, which often have lower gene density and may therefore be less subject to linked selection (Kulikova et al. 2001; Paape et al. 2013). To verify that this is a possible explanation, we computed gene densities in each window using gene models in Mt4.0 from `jcvi.org` (Tang et al. 2014), which are shown juxtaposed with the first MDS coordinate of each window is shown in Figure 8, and are significantly correlated, as shown in Figure 9. (Values shown are the number of start and end positions of each predicted mRNA transcript, divided by two, assigned to the nearest window.) Other genomic features covary with gene density, however, so we cannot rule out alternative hypotheses.

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

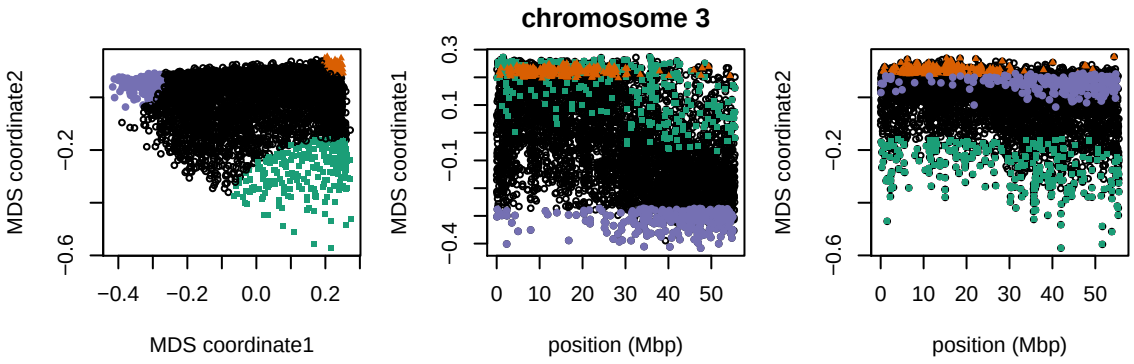

Figure 6: MDS visualization of population structure on *M. truncatula* chromosome 3. Each point in the plot represents a window. Plots for remaining chromosomes are shown in Supplemental Figure S6.

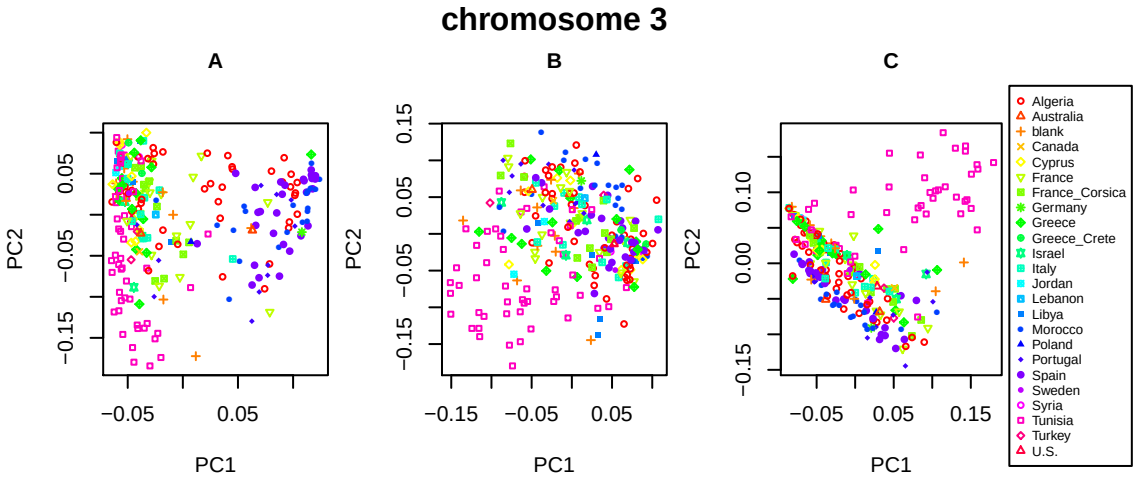

Figure 7: PCA plots for the sets of genomic windows colored (A) green, (B) orange, and (C) purple in Figure 6. Each point corresponds to a sample, colored by country of origin. Plots for remaining chromosomes are shown in Supplemental Figure S7.

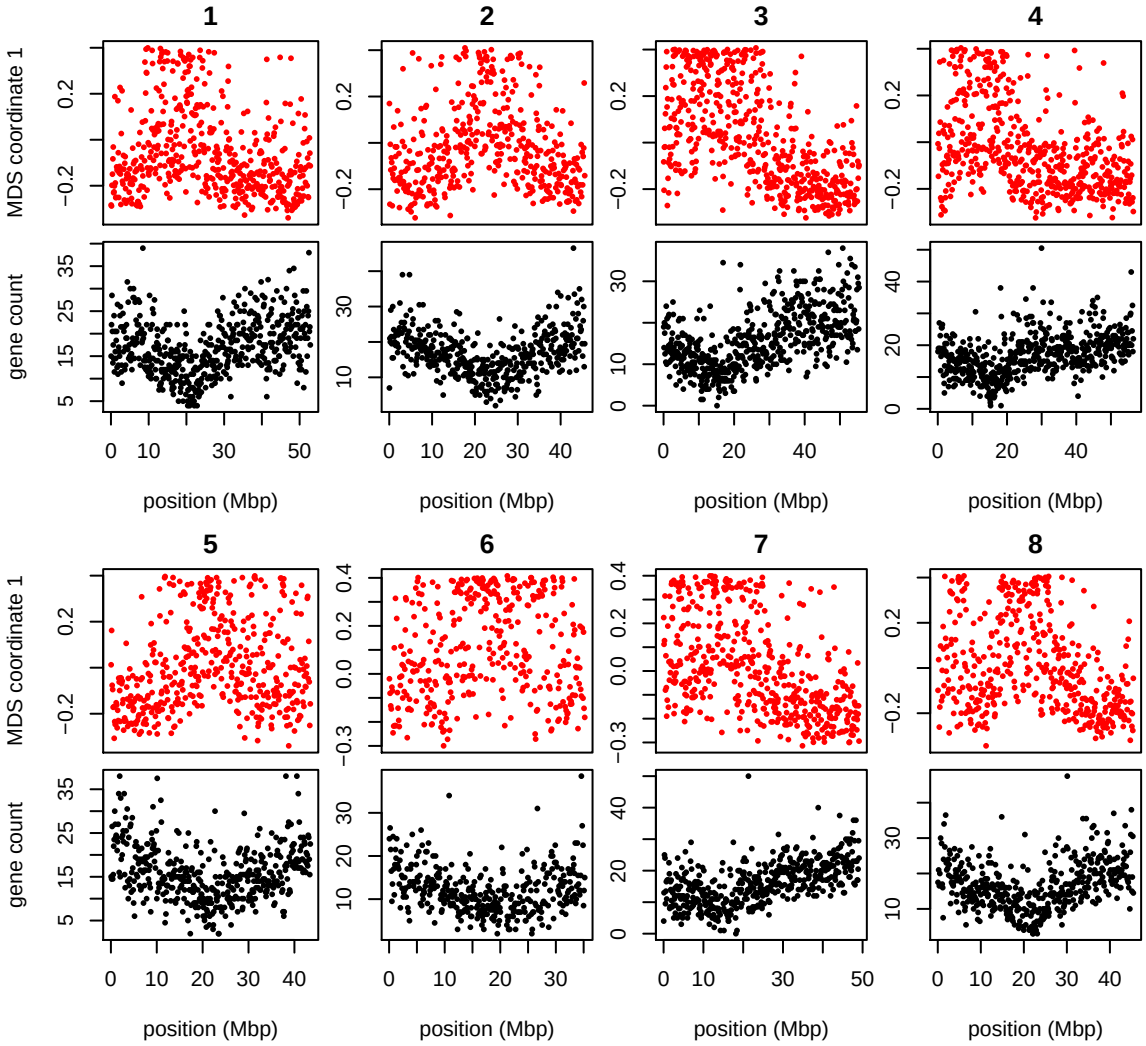

Figure 8: MDS coordinate and gene density for each window in the *Medicago* genome, for chromosomes 1–8 (numbered above each pair of figures). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene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

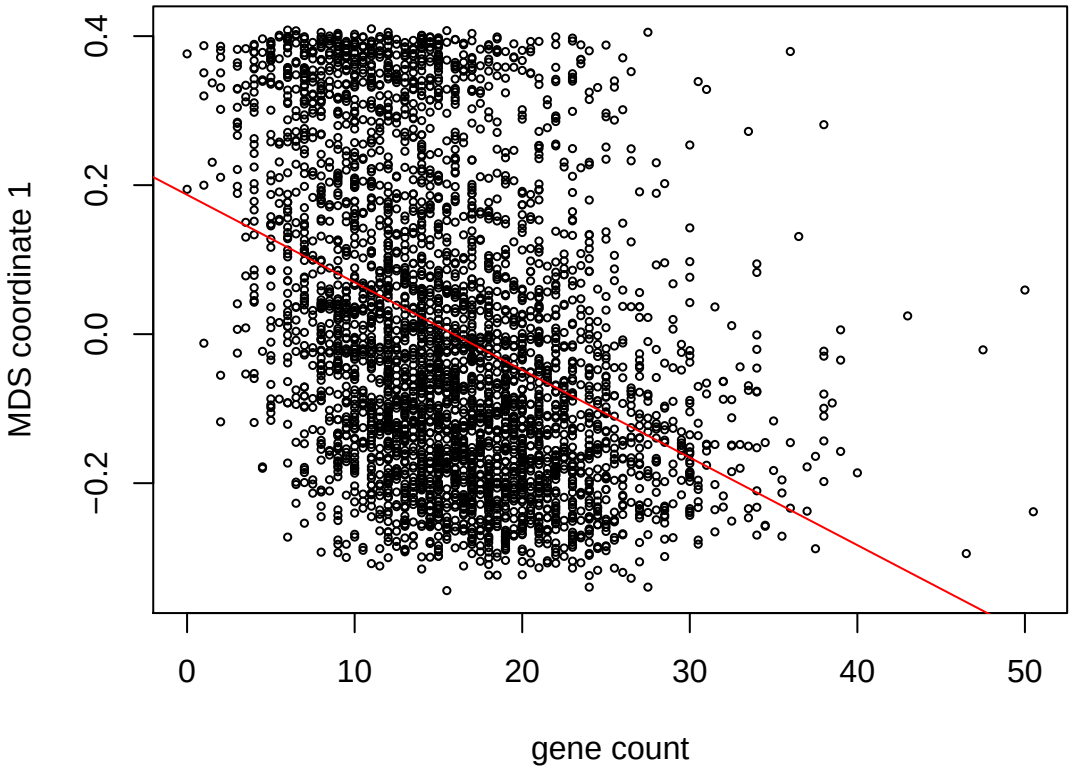

Figure 9: First MDS coordinate against gene density for all 8 chromosomes of *M. truncatula*. The first MDS coordinate is significantly correlated with gene count ($r = 0.149$, $p = 2.2 \times 10^{-16}$).

4 Discussion

Our investigations have found a remarkable amount of variation in population structure across the genomes of three diverse species, revealing distinct biological processes driving this variation in each species. More investigation, particularly on more species and datasets, will help to uncover which patterns are generalizable.

With growing appreciation of the heterogeneous effects of selection across the genome, especially the importance of adaptive introgression, hybrid speciation (Brandvain et al. 2014; Fitzpatrick et al. 2010; Hufford et al. 2013; Pool 2015; Staubach et al. 2012), local adaptation (Lenormand 2002; Wang and Bradburd 2014), and inversion polymorphisms (Kirkpatrick 2010; Kirkpatrick and Barrett 2015), local PCA may prove to be a useful exploratory tool to discover important genomic features.

Confounding in GWAS So-called cryptic relatedness between samples has been one of the major sources of confounding in genome-wide association studies (GWAS) and so methods must account for it by modeling population structure or kinship (Astle and Balding 2009; Yang et al. 2014). Since population structure is not constant along the genome, this could in principle lead to an inflation of false positives in parts of the genome with stronger population structure than the genome-wide average. Fortunately, in our human dataset this does not seem likely to have a strong effect: most variation is due to small, independent regions, possibly primarily inversions, and so may not have a major effect on GWAS. In the other species we examined, particularly *Drosophila melanogaster*, treating population structure as a single quantity could be severely misleading.

The effect of selection How can we expect selection to affect population structure? PCA summarizes patterns in kinship found in the genetic covariance matrix, which is an

average across locus-specific genealogies: if individuals are closer in the genealogies of a given genomic region, they tend to be closer in the PCA maps for that region. Strong selective sweeps of beneficial alleles can lead to relatively long genomic regions characterized by short genealogical trees (Garud et al. 2013; Przeworski et al. 2005). A genomic region with many targets for selection experiences background selection and/or recurrent selective sweeps (Coop and Ralph 2012; Stephan et al. 1992), which would tend to shorten genealogical trees in the region, similar to a reduction in effective population size (Hudson and Kaplan 1995; Sattath et al. 2011). On the other hand, balancing selection leads to very deep trees (Gao et al. 2015), and population structure locally describes which individuals have which alleles rather than geographical proximity. Finally, since recombination is suppressed between opposite orientations of a chromosomal inversion near its breakpoints, genealogies in these regions separate samples carrying the two orientations of the inversion.

Many of the effects listed above (such as single selective sweeps or inversions) are not expected to have similar effects on population structure in different regions of the genome because of randomness in which samples end up in which group; but if these coincide with a region of reduced recombination (or an inversion), these could drive major patterns of variation. The more subtle effects of genome-wide linked selection could be shared across large regions due to large-scale variation in gene density (as by background selection or local adaptation at many genes).

Parameter choices There are several choices in the method that may affect the results. As with whole-genome PCA, the choice of samples is important, as variation not strongly represented in the sample will not be discovered. The effects of strongly imbalanced sampling schemes are often corrected by dropping samples in overrepresented groups; but downweighting may be a better option that does not discard data (and here we present a method to do this). Next, the choice of window size may be important, although in our

applications results were not sensitive to this, indicating that we can see variation on a sufficiently fine scale. Finally, which collections of genomic regions are compared to each other (steps 3 and 4 in Figure 1), along with the method used to discover common structure, will affect results. We used MDS, applied to either each chromosome separately or to the entire genome; for instance, human inversions are clearly visible as outliers when compared to the rest of their chromosome, but genome-wide, their signal is obscured by the numerous other signals of comparable strength.

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

More generally, there are many possible methods to discover common structure in different parts of the genome: in this work we use PCA to summarize population structure, but other methods, such as STRUCTURE (Falush et al. 2003), SPA (Yang et al. 2012), SpaceMix (Bradburd et al. 2016), or other matrix factorization methods (Engelhardt and Stephens 2010) may highlight different sources of variation in the data. It is also possible that other methods for measuring dissimilarity between windows' covariance matrices or for summarizing the matrix of pairwise distances between windows would lead to different insights.

Finally, our software allows different choices for how many PCs to use in approximating structure of each window (k in equation 1), and how many MDS coordinates to use when describing the distance matrix between windows, but in our exploration, changing these has not produced dramatically different results. These are all part of more general techniques

in dimension reduction and high-dimensional data visualization; we encourage the user to experiment.

Chromosomal inversions A major driver of variation in population structure in two datasets we examined are inversions. This may be common, but the example of *Medicago truncatula* shows that polymorphic inversions are not ubiquitous. PCA has been proposed as a method for discovering inversions (Ma and Amos 2012); however, the signal left by inversions likely cannot be distinguished from long haplotypes under balancing selection or simply regions of reduced recombination without additional lines of evidence. However, in many applications, inversions are a nuisance. For instance, SMARTPCA (Patterson et al. 2006) reduces their effect on PCA plots by regressing out the effect of linked SNPs on each other. It would be interesting to see if somehow removing the effects of inversions in the *Drosophila melanogaster* or human datasets would produce a pattern similar to that seen in *Medicago truncatula*.

Acknowledgements

We are indebted to John Pool, Russ Corbett-Detig, Matilde Cordeiro, and Peter Chang for assistence with obtaining data and interpreting results (especially inversion status of *D. melanogaster* samples). Thanks also go to Yaniv Brandvain and Barbara Engelhardt for helpful comments and for encouraging the project.

References

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

William Astle and David J. Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 11 2009. doi: 10.1214/09-STS307. URL http://dx.doi.org/10.1214/09-STS307.

Nicholas H. Barton. Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, 355(1403): 1553–1562, November 2000. doi: 10.1098/rstb.2000.0716. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1692896/.

Gideon S. Bradburd, Peter L. Ralph, and Graham M. Coop. A spatial framework for understanding population structure and admixture. *PLoS Genet*, 12(1):1–38, 01 2016. doi: 10.1371/journal.pgen.1005703. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1005703.

Yaniv Brandvain, Amanda M. Kenney, Lex Flagel, Graham Coop, and Andrea L. Sweigart. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genet*, 10(6):e1004410, 06 2014. doi: 10.1371/journal.pgen.1004410. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1004410.

Frank M.T.A. Busing, Erik Meijer, and Rien Van Der Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9(1):3–8, 1999. ISSN 0960-3174. doi: 10.1023/A:1008800423698. URL http://dx.doi.org/10.1023/A%3A1008800423698.

Brian Charlesworth. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1):5–22, January 2012. doi: 10.1534/genetics.111.134288. URL http://www.ncbi.nlm.nih.gov/pubmed/22219506.

Brian Charlesworth, Deborah Charlesworth, and Nicholas H. Barton. The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):99–125, 2003. doi: 10.1146/annurev.ecolsys.34.011802.132359.

URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.34.011802.132359.

Graham Coop and Peter Ralph. Patterns of neutral diversity under general models of selective sweeps. *Genetics*, 192(1):205–224, September 2012. doi: 10.1534/genetics.112.141861. URL http://www.ncbi.nlm.nih.gov/pubmed/22714413.

Russell B Corbett-Detig and Daniel L Hartl. Population genomics of inversion polymorphisms in drosophila melanogaster. *PLoS Genet*, 8(12):e1003056, 2012.

Russell B. Corbett-Detig, Charis Cardeno, and Charles H. Langley. Sequence-based detection and breakpoint assembly of polymorphic inversions. *Genetics*, 192(1):131–137, 2012. URL http://www.genetics.org/content/192/1/131.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, 1982. doi: 10.1137/1.9781611970319. URL http://epubs.siam.org/doi/abs/10.1137/1.9781611970319.

Hans Ellegren, Linnea Smeds, Reto Burri, Pall I. Olason, Niclas Backstrom, Takeshi Kawakami, Axel Kunstner, Hannu Makinen, Krystyna Nadachowska-Brzyska, Anna Qvarnstrom, Severin Uebbing, and Jochen B. W. Wolf. The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491(7426):756–760, November 2012. ISSN 00280836. doi: 10.1038/nature11584. URL http://dx.doi.org/10.1038/nature11584.

Barbara E. Engelhardt and Matthew Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet*, 6(9), September 2010. doi: 10.1371/journal.pgen.1001117. URL http://www.ncbi.nlm.nih.gov/pubmed/20862358.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, August 2003. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462648/.

B M Fitzpatrick, J R Johnson, D K Kump, J J Smith, S R Voss, and H B Shaffer. Rapid spread of invasive genes into a threatened native species. *Proc Natl Acad Sci U S A*, 107(8):3606–3610, February 2010. doi: 10.1073/pnas.0911802107. URL http://www.ncbi.nlm.nih.gov/pubmed/20133596.

Ziyue Gao, Molly Przeworski, and Guy Sella. Footprints of ancient balanced polymorphisms in genetic variation data. *Evolution*, 69(2):431–446, 2015. doi: 10.1111/evo.12567. URL http://onlinelibrary.wiley.com/doi/10.1111/evo.12567/abstract.

Nandita R. Garud, Philipp W. Messer, Erkan O. Buzbas, and Dmitri A. Petrov. Soft selective sweeps are the primary mode of recent adaptation in *Drosophila melanogaster*, 2013. URL http://arxiv.org/abs/1303.0906. cite arxiv:1303.0906.

Rafael F. Guerrero, François Rousset, and Mark Kirkpatrick. Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1587):430–438, 2011. ISSN 0962-8436. doi: 10.1098/rstb.2011.0246. URL http://rstb.royalsocietypublishing.org/content/367/1587/430.

R. R. Hudson and N. L. Kaplan. Deleterious background selection with recombination. *Genetics*, 141(4):1605–1617, December 1995. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1206891/.

Matthew B. Hufford, Pesach Lubinksy, Tanja Pyhjrvi, Michael T. Devengenzo, Norman C.

Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgression in maize. *PLoS Genet*, 9(5):e1003477, 05 2013. doi: 10.1371/journal.pgen.1003477. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1003477.

N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, July 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.7.1493. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6795533.

Mark Kirkpatrick. How and why chromosome inversions evolve. *PLoS Biol*, 8(9), 2010. doi: 10.1371/journal.pbio.1000501. URL http://www.ncbi.nlm.nih.gov/pubmed/20927412.

Mark Kirkpatrick and Brian Barrett. Chromosome inversions, adaptive cassettes and the evolution of species' ranges. *Molecular Ecology*, 2015. ISSN 1365-294X. doi: 10.1111/mec.13074. URL http://dx.doi.org/10.1111/mec.13074.

Olga Kulikova, Gustavo Gualtieri, René Geurts, Dong-Jin Kim, Douglas Cook, Thierry Huguet, J Hans De Jong, Paul F Fransz, and Ton Bisseling. Integration of the FISH pachytene and genetic maps of *Medicago truncatula*. *The Plant Journal*, 27(1):49–58, 2001.

Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-Detig, Kristian A Stevens, Charles H Langley, and John E Pool. The Drosophila genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241, 2015.

C H Langley, K Stevens, C Cardeno, Y C Lee, D R Schrider, J E Pool, S A Langley, C Suarez, R B Corbett-Detig, B Kolaczkowski, S Fang, P M Nista, A K Holloway, A D

Kern, C N Dewey, Y S Song, M W Hahn, and D J Begun. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, 192(2):533–598, October 2012. doi: 10.1534/genetics.112.142018. URL http://www.ncbi.nlm.nih.gov/pubmed/22673804.

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.

J Ma and C I Amos. Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One*, 7(7), 2012. doi: 10.1371/journal.pone.0040224. URL http://www.ncbi.nlm.nih.gov/pubmed/22808122.

José V Manjón, Pierrick Coupé, Luis Concha, Antonio Buades, D Louis Collins, and Montserrat Robles. Diffusion weighted image denoising using overcomplete local PCA. *PloS one*, 8(9):e73021, 2013.

J Maynard Smith and J Haigh. The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35, February 1974. URL http://www.ncbi.nlm.nih.gov/pubmed/4407212.

Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genet*, 5(10):e1000686, 2009.

P Menozzi, A Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792, September 1978. URL http://www.ncbi.nlm.nih.gov/pubmed/356262.

N J Nadeau, A Whibley, R T Jones, J W Davey, K K Dasmahapatra, S W Baxter, M A Quail, M Joron, R H ffrench Constant, M L Blaxter, J Mallet, and C D Jiggins.

Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci*, 367(1587):343–353, February 2012. doi: 10.1098/rstb.2011.0198. URL `http://www.ncbi.nlm.nih.gov/pubmed/22201164`.

Richard A. Neher. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44:195–215, November 2013. doi: 10.1146/annurev-ecolsys-110512-135920. URL `http://arxiv.org/abs/1302.1148`.

M R Nelson, K Bryc, K S King, A Indap, A R Boyko, J Novembre, L P Briley, Y Maruyama, D M Waterworth, G Waeber, P Vollenweider, J R Oksenberg, S L Hauser, H A Stirnadel, J S Kooner, J C Chambers, B Jones, V Mooser, C D Bustamante, A D Roses, D K Burns, M G Ehm, and E H Lai. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*, 83(3): 347–358, September 2008. doi: 10.1016/j.ajhg.2008.08.005. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556436/?tool=pubmed`.

John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

Timothy Paape, Thomas Bataillon, Peng Zhou, Tom JY Kono, Roman Briskine, Nevin D Young, and Peter Tiffin. Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. *Molecular ecology*, 22(13):3525–3538, 2013.

Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 12 2006. doi: 10.1371/journal.pgen.0020190. URL http://dx.plos.org/10.1371%2Fjournal.pgen.0020190.

J B Pease and M W Hahn. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, 67(8):2376–2384, August 2013. doi: 10.1111/evo.12118. URL http://www.ncbi.nlm.nih.gov/pubmed/23888858.

John E Pool. The mosaic ancestry of the Drosophila Genetic Reference Panel and the *D. melanogaster* reference genome reveals a network of epistatic fitness interactions. *Molecular Biology and Evolution*, 32(12):3236–3251, 2015. doi: 10.1101/014837. URL http://mbe.oxfordjournals.org/content/32/12/3236.abstract.

John E. Pool, Russell B. Corbett-Detig, Ryuichi P. Sugino, Kristian A. Stevens, Charis M. Cardeno, Marc W. Crepeau, Pablo Duchen, J. J. Emerson, Perot Saelao, David J. Begun, and Charles H. Langley. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet*, 8(12):1–24, 12 2012. doi: 10.1371/journal.pgen.1003080. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1003080.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

M Przeworski, G Coop, and J D Wall. The signature of positive selection on standing genetic variation. *Evolution*, 59(11):2312–2323, November 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16396172.

Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems*,

2016. URL https://CRAN.R-project.org/package=RSpectra. R package version 0.11-0.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2323. URL http://science.sciencemag.org/content/290/5500/2323.

S Sattath, E Elyashiv, O Kolodny, Y Rinott, and G Sella. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. *PLoS Genet*, 7(2), 2011. doi: 10.1371/journal.pgen.1001302. URL http://www.ncbi.nlm.nih.gov/pubmed/21347283.

Fabian Staubach, Anna Lorenc, Philipp W. Messer, Kun Tang, Dmitri A. Petrov, and Diethard Tautz. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genet*, 8(8):e1002891, 08 2012. doi: 10.1371/journal.pgen.1002891. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1002891.

Wolfgang Stephan, Thomas H.E. Wiehe, and Marcus W. Lenz. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoretical Population Biology*, 41(2):237 – 254, 1992. ISSN 0040-5809. doi: http://dx.doi.org/10.1016/0040-5809(92)90045-U. URL http://www.sciencedirect.com/science/article/pii/004058099290045U.

Haibao Tang, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo Zhou, Laurent Gentzbittel, Kevin L Childs, Mark Yandell, Heidrun Gundlach, et al. An improved genome release (version mt4. 0) for the model legume *Medicago truncatula*. *BMC genomics*, 15(1):1, 2014.

Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10:66–71, 2009.

Benjamin Vernot and Joshua M. Akey. Resurrecting surviving neandertal lineages from modern human genomes. *Science*, 2014. doi: 10.1126/science.1245938. URL `http://www.sciencemag.org/content/early/2014/01/28/science.1245938.abstract`.

Ian J. Wang and Gideon S. Bradburd. Isolation by environment. *Molecular Ecology*, 23 (23):5649–5662, 2014. ISSN 1365-294X. doi: 10.1111/mec.12938. URL `http://dx.doi.org/10.1111/mec.12938`.

Andreas Weingessel and Kurt Hornik. Local PCA algorithms. *Neural Networks, IEEE Transactions on*, 11(6):1242–1250, 2000.

Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46(2):100–106, February 2014. ISSN 10614036. URL `http://dx.doi.org/10.1038/ng.2876`.

W Y Yang, J Novembre, E Eskin, and E Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*, 44(6):725–731, June 2012. doi: 10.1038/ng.2285. URL `http://www.ncbi.nlm.nih.gov/pubmed/22610118`.

A Weighted PCA

Principal components analysis can be thought of as finding a good low-dimensional matrix factorization (Engelhardt and Stephens 2010) that well-approximates the original data in the least-squares sense: if C is the $N \times N$ genetic covariance matrix, then to find the top k

principal components, we find an orthogonal $N \times k$ matrix U, and a $k \times k$ diagonal matrix Λ with diagonal entries $\Lambda_{ii} = \lambda_i$ to minimize

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

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

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

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

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

and now for convenince we require U to be orthogonal in $\ell_2(w)$, i.e., that $U^T W U = I$. We then would choose w to give roughly equal weight to each *population*, instead of each individual. We have used with good results the weightings $w_i = 1/\max(10, n_i)$, where

n_i is, if there are discrete populations, the number of samples in the same population as sample i; or, for continuously sampled individuals, the number of samples within a certain distance of sample i.

To solve (4), let λ and V denote the top k eigenvalues and eigenvectors of $W^{1/2}CW^{1/2}$, 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 $U = W^{-1/2}V$ then $U^TWU = V^TV = I$, and

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

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

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

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

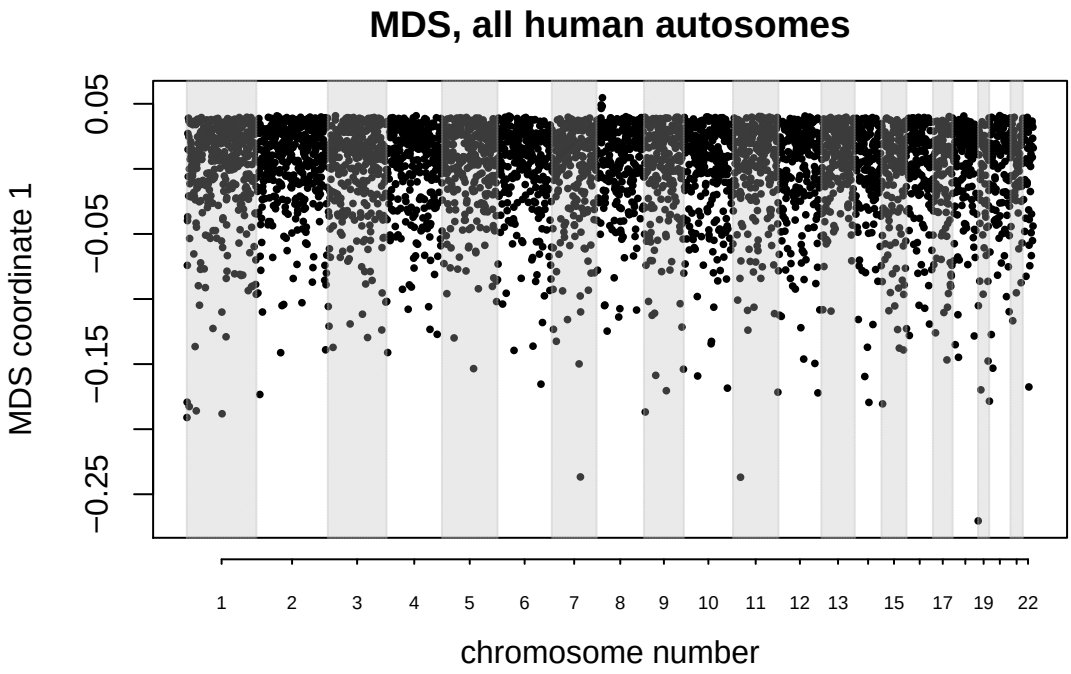

Figure S1: MDS visualization of variation in population structure amongst windows across *all* human autosomes simultaneously. The small group of windows with positive outlying MDS values lie around the inversion at 8p23.

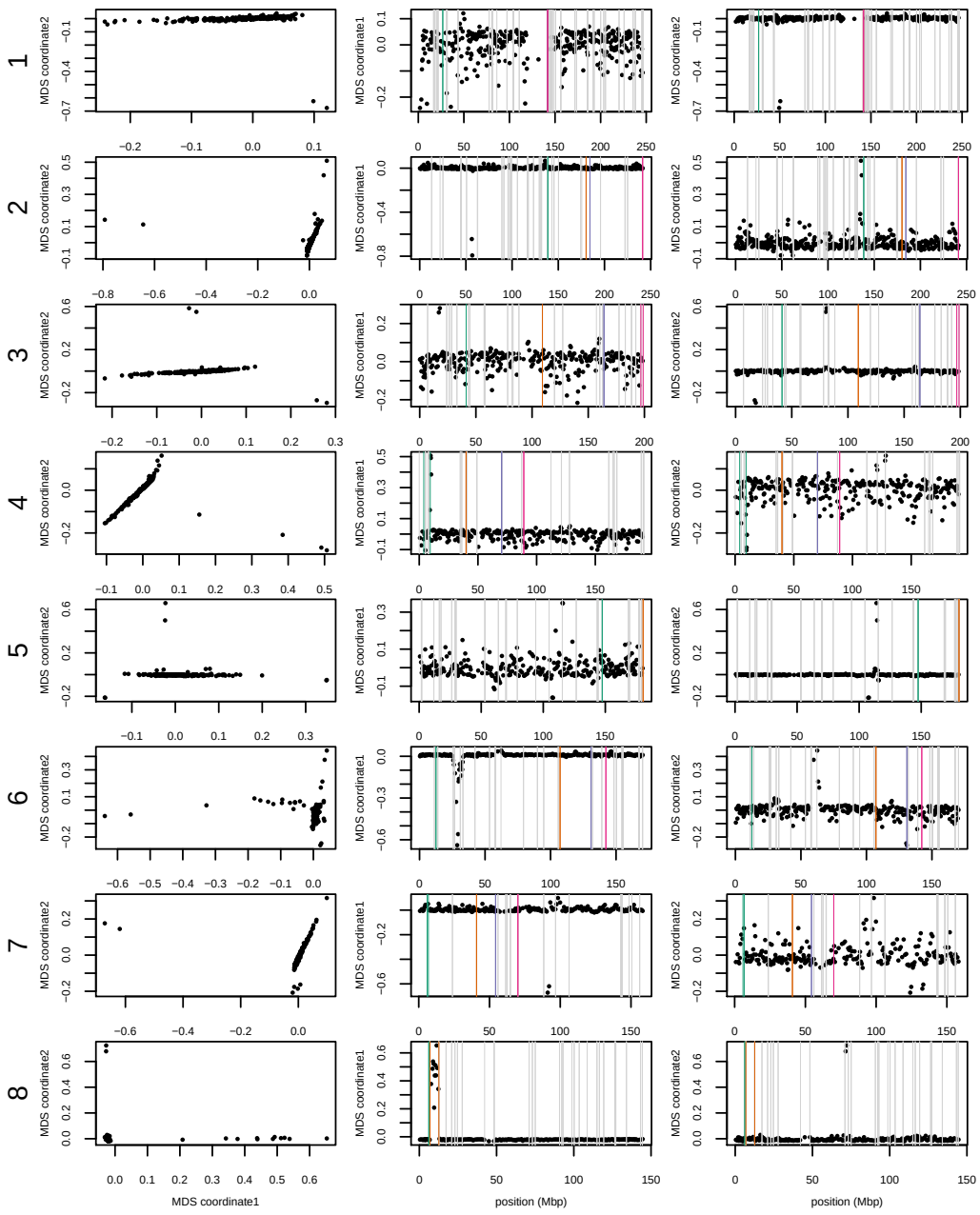

Figure S2: MDS plots for human chromosome 1-8. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosomes. Colorful vertical lines show the breakpoints of known valid inversions, while grey vertical lines show the breakpoints of predicted inversions.

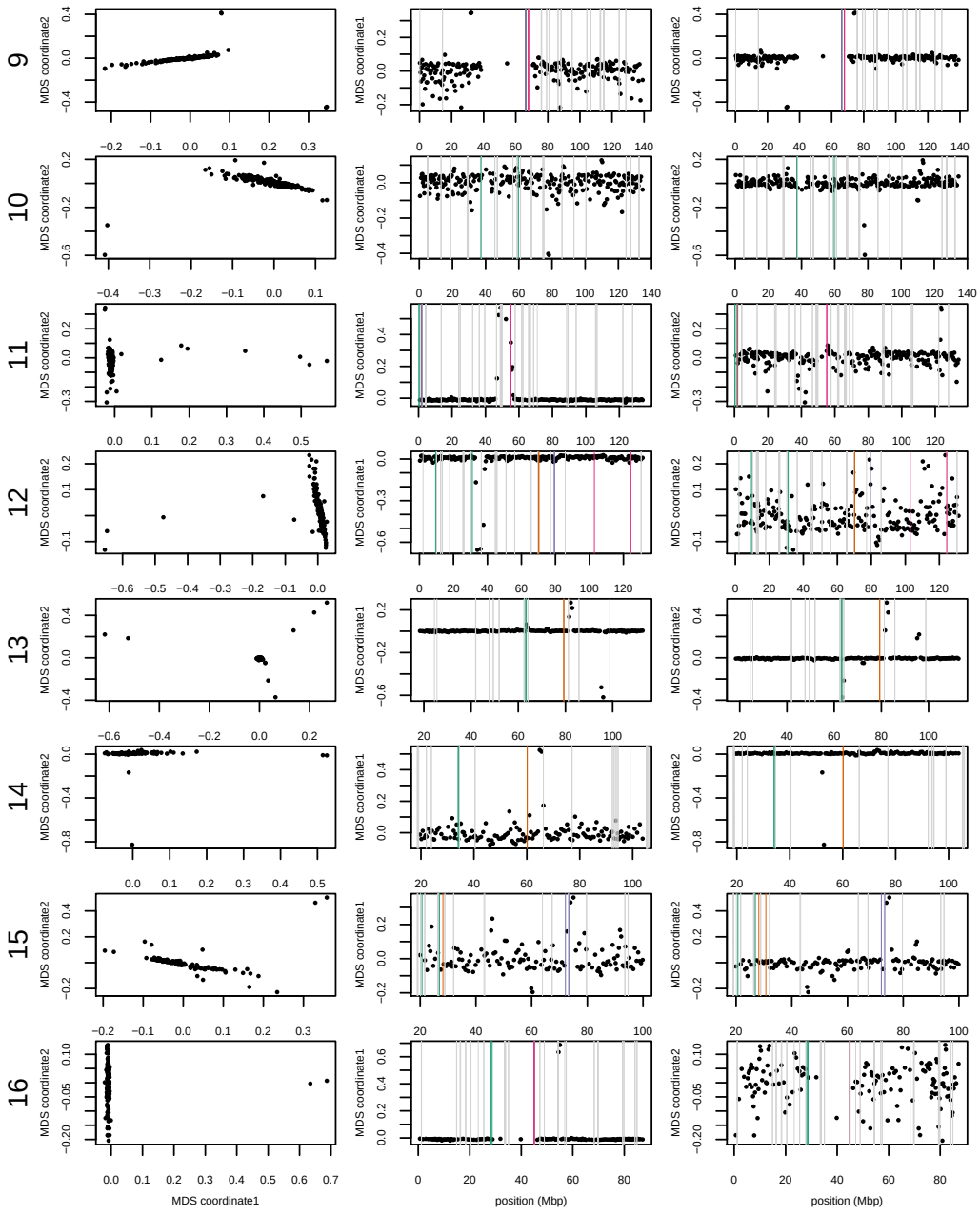

Figure S3: MDS plots for human chromosome 9-16.

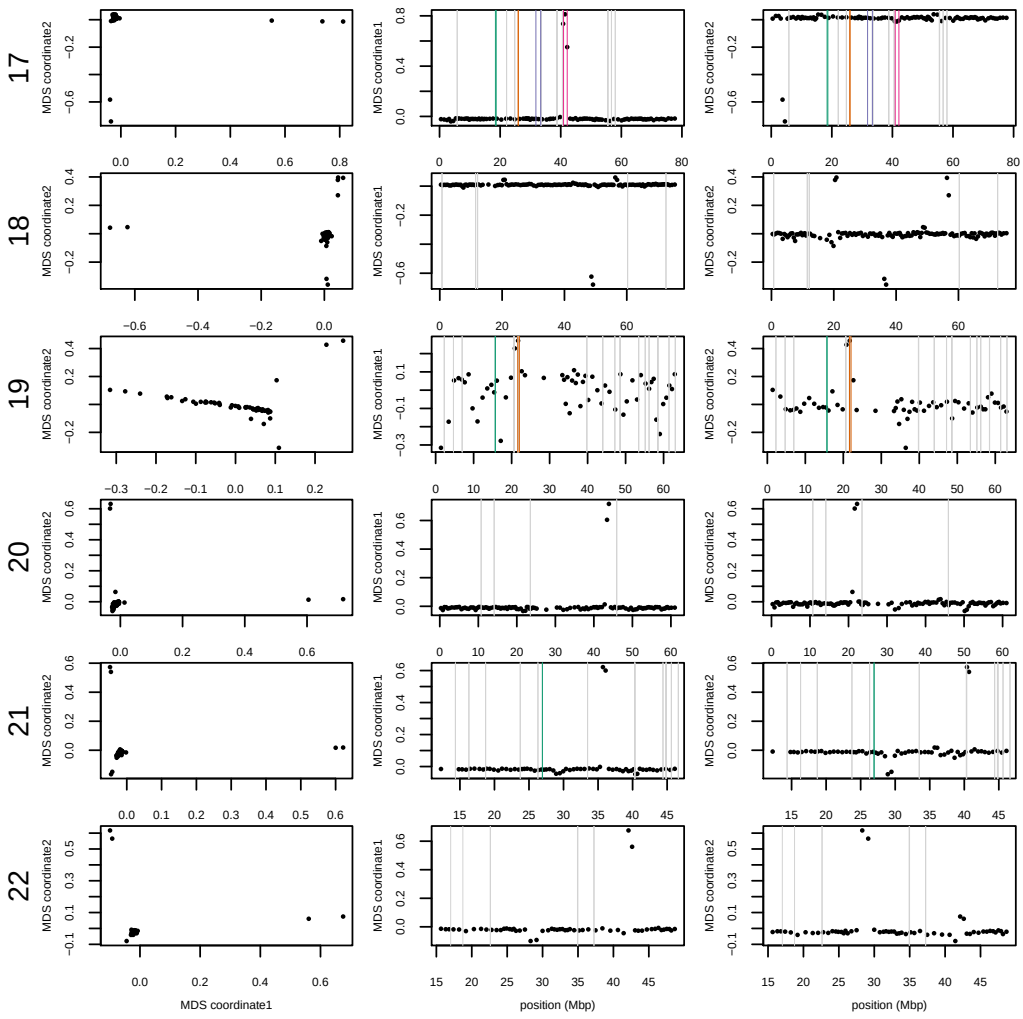

Figure S4: MDS plots for human chromosome 17-22.

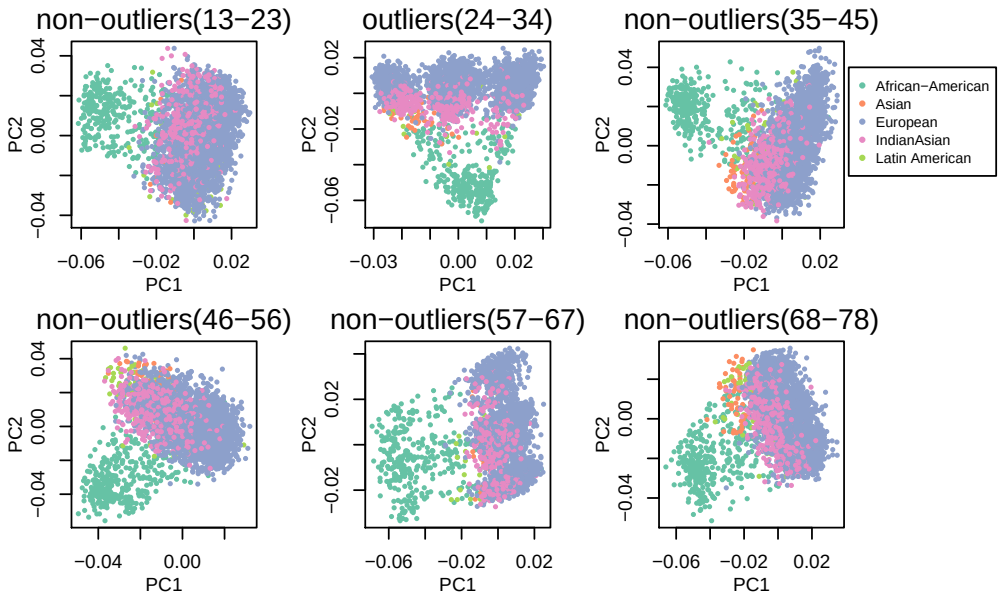

Figure S5: PCA plots comparison for outlier windows and non-outlier windows.

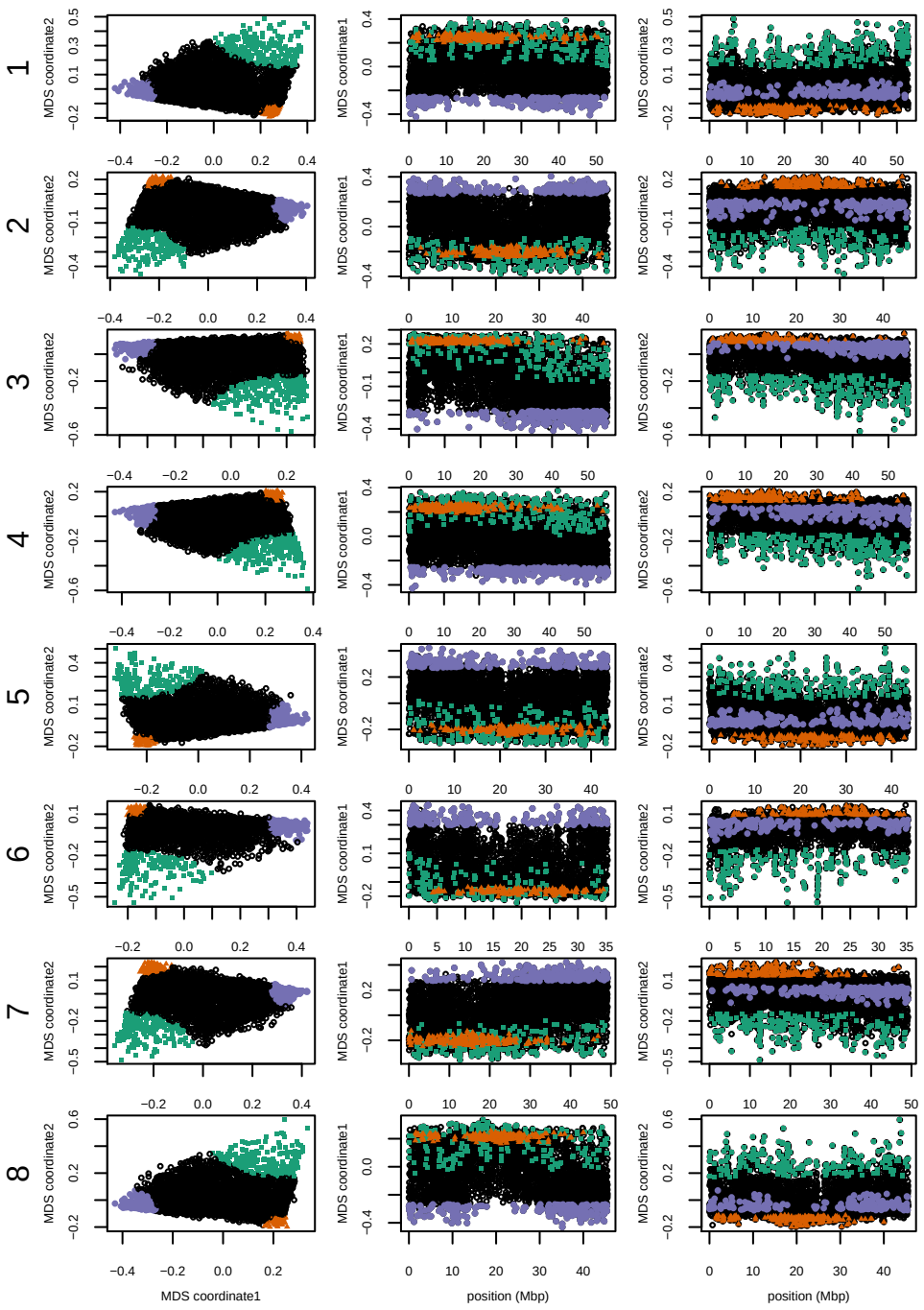

Figure S6: MDS visualizations of population structure for all 8 chromosomes of the *Medicago truncatula* data.

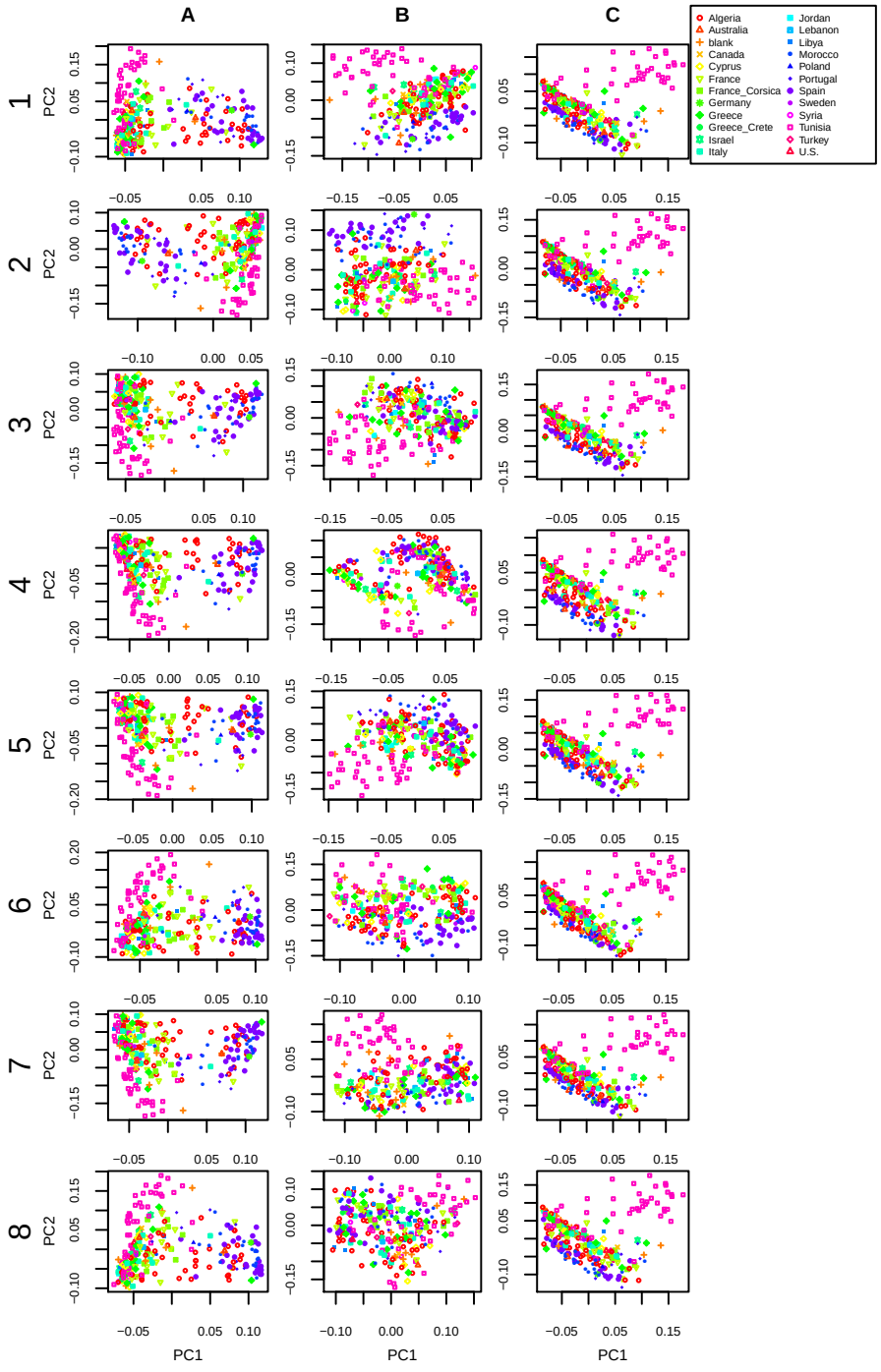

Figure S7: PCA plots for regions colored in Figure S6 on all 8 chromosomes of *Medicago truncatula*: (A) green, (B) orange, and (C) purple.

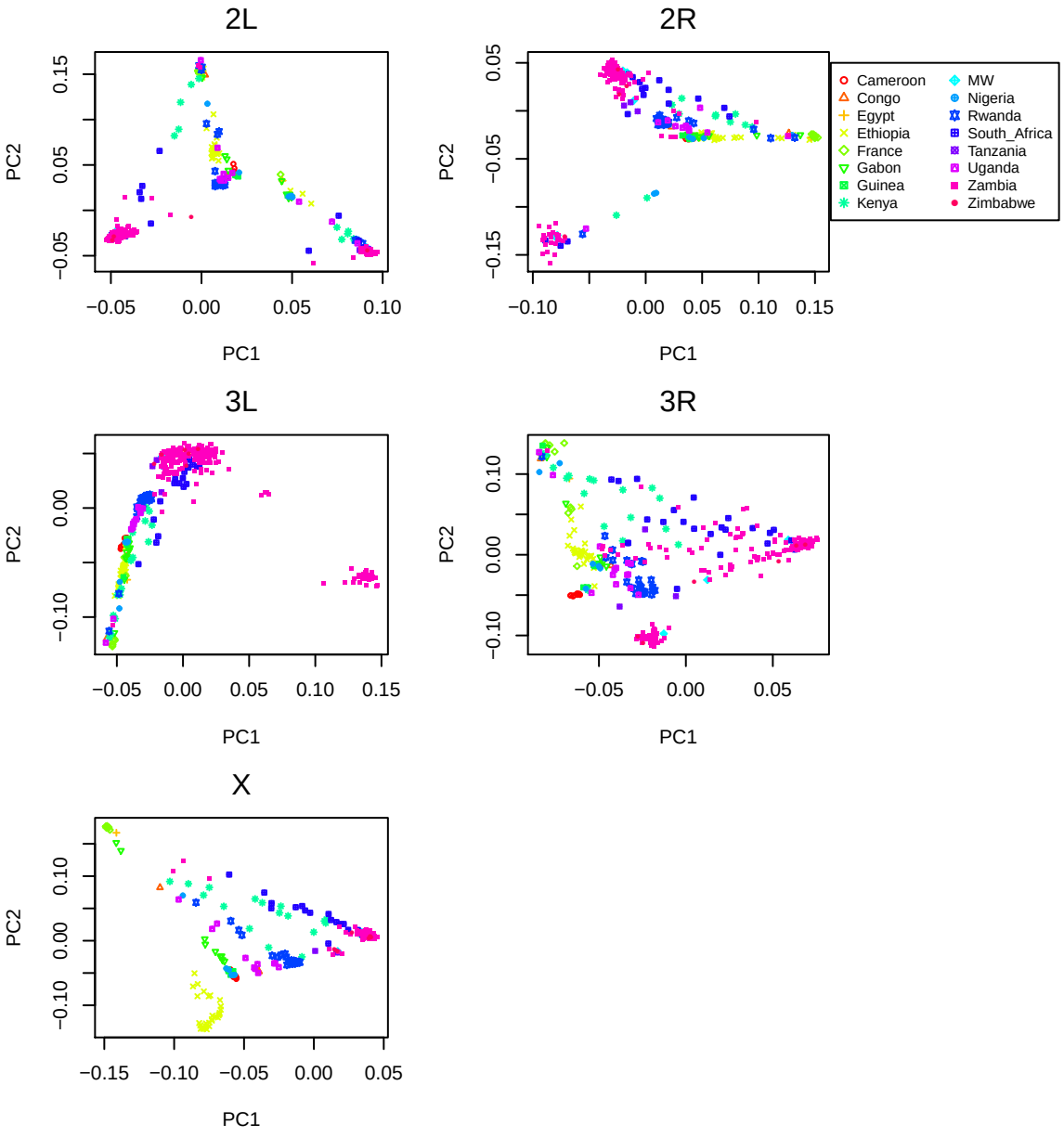

Figure S8: PCA plots for chromosome arms 2L, 2R, 3L, 3R and X of the *Drosophila melanogaster* dataset.

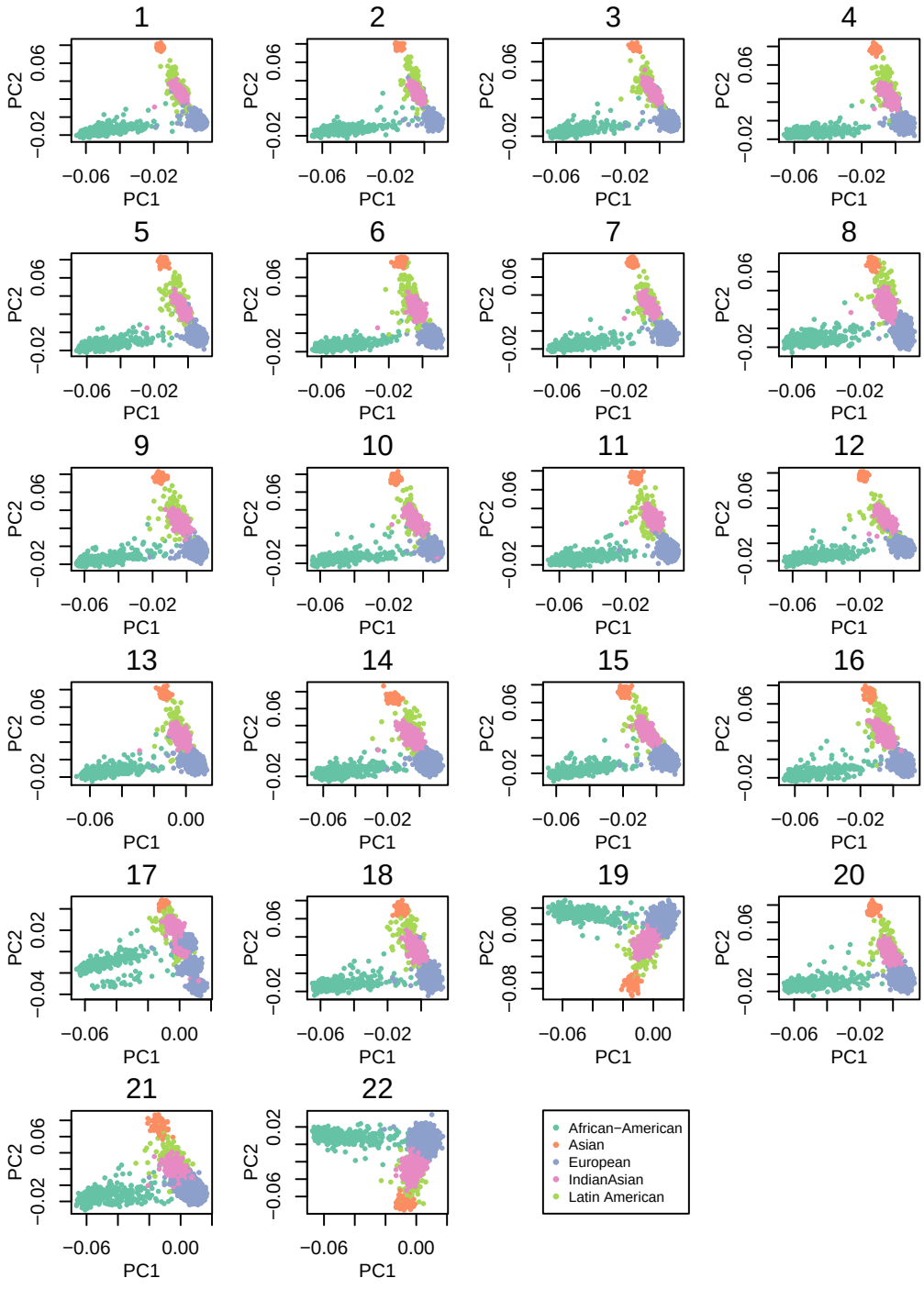

Figure S9: PCA plots for all 22 huan autosomes from the POPRES data.

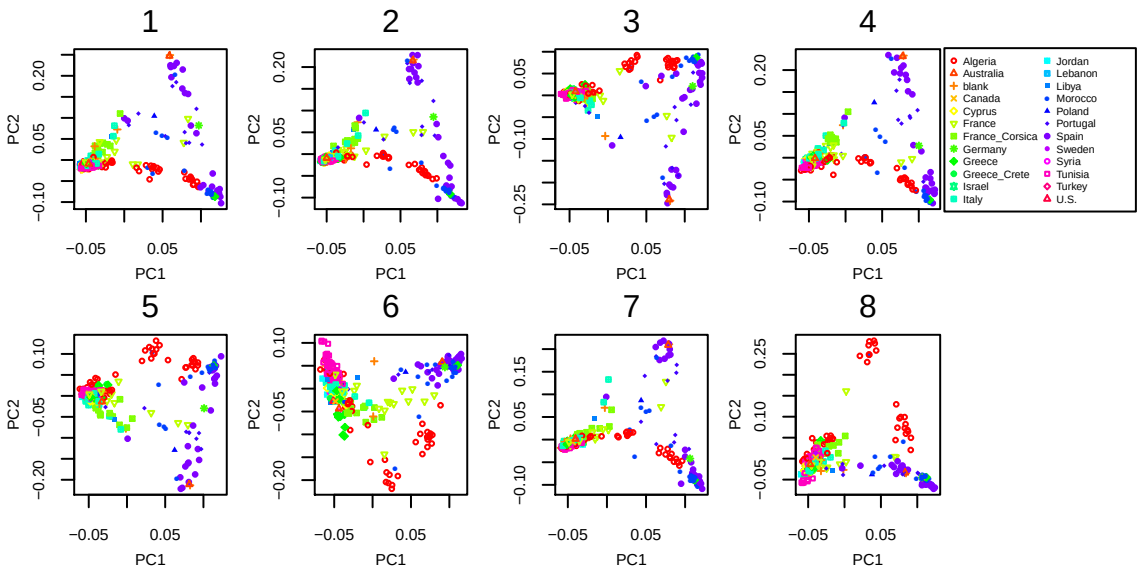

Figure S10: PCA plots for all 8 chromosomes in the *Medicago truncatula* dataset.