Local PCA Shows How Population Structure Differs Along the Genome

Han Li, Peter Ralph

June 3, 2016

Abstract

Dimension reduction techniques, such as principal components, are often used to discover and display large-scale structure in genomic datasets found in the patterns of kinship between the genotyped individuals, and to control for the confounding effects of population structure in genome-wide association studies. The genome-wide mean kinship this uses is an average of the relationships across all locus-specific genealogical trees. However, many biological factors, including linked selection, can systematically skew patterns of kinship over intermediate genomic scales. We show how to use principal components analysis (PCA) to describe this meso-scale variation in kinship, and apply the method to genomic data from three species. In each species we find substantial variation 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 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 a useful addition to the exploratory toolbox of population

genomics.

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

Add whole-genome PCA plots for comparison (unless they look just like one of the three corners; in which case say so.

1 Introduction

The phrase "population structure" refers to reduced gene flow between subpopulations, often because of geographical isolation. 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, paradoxically, the population structure of a species depends on which part of the genome is being examined.

Population structure leads to systematic patterns in genome-wide mean kinship, and so it is said that visualizations of kinship depict population structure, rather than the *effects* of population structure (which would be more accurate). The *kinship coefficient* for a pair of individuals gives the expected proportion of their genome that the two have inherited identically by descent; for a single individual it is the inbreeding coefficient. However, as Wright (1949) wrote, "It has probably occurred to the reader that the coefficient of inbreeding may mean very different things in different cases." The kinship coefficient originally referred to the expected probability of coinheritance within a given pedigree; while modern applications to "unrelated" individuals use a genetic covariance matrix to estimate the *realized* proportion of the genome coinherited from sufficiently recent ancestors.

Realized kinship summarizes the shapes of the genealogical that relate the samples 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), and therefore patterns of kinship. Indeed, chromosome-scale variation in diversity and divergence has been observed in many species (e.g. 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 what the major axes of variation are.

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 directly related to the underlying genealogical history of the samples, such as time to most recent common ancestor and migration rate between populations (McVean 2009; Novembre and Stephens 2008). These patterns are often geographical, producing "maps" of population structure that reflect the samples' geographic origin distorted by rates of gene flow (Novembre et al. 2008), although other patterns emerge if recent migration or nongeographic kinship patterns are important (Astle and Balding 2009).

As reviewed by Astle and Balding (2009), modeling "background" kinship between samples is essential to successful genome-wide association studies, and PCA has often been used for stratification correction (Price et al. 2006), and so investigating variation

in kinship along the genome can help us to have a better understanding of the relation between genome structure and population structure, possibly leading to more powerful methods for association studies.

To investigate how population structure varies along the genome in several datasets, we cut each genome into windows (with hundreds to thousands of SNPs in each), applied PCA to each window, and visualized the major ways that population structure, as summarized by PCA, varies among windows. whole genome sequencing data for *Drosophila*. To quantify similarity of population structure between windows, we constructed for each window an approximate, scaled covariance matrix based on the first few principal components, and measured the pairwise Euclidean distance between those matrices. We then use multidimensional scaling to visualize the relationships between windows, which reduces the pairwise distance matrix to lower dimension while preserving the distance information between windows as well as possible (Borg and Groenen 2005).

Each species showed distinct patterns, reflecting differences in their biology; before presenting results, it may help to have a prior idea of what factors are expected to affect population structure, and how. 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 tends to be more close 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 generally 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. 2014), 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. In the resulting extended block, population structure shows two (for haploids) or three (for diploids) clusters (indeed, (Ma and Amos 2012) has proposed using trimodality of local PCA plots as a way to identify inversions). We will look for strong variation in population structure shared across large regions of the genome. 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 (such as 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).

We chose PCA to summarize population structure, but other methods, such as STRUCTURE (Falush et al. 2003), SPA (Yang et al. 2012), SpaceMix (Bradburd et al. 2015), or other matrix factorization methods (Engelhardt and Stephens 2010) would highlight different sources of variation in the data.

Finally, 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 use of the term 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. Referring to 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); as usual for genetic data we converted the data 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.

Human: We used genomic data from the entire POPRES dataset (Nelson et al. 2008), which has array-derived genotype information for 447,267 SNPs across the autosomes of 3,965 samples in total: 346 African-Americas, 73 Asians, 3,187 Europeans and 359 Indian Asians. (We excluded the sex chromosomes and the mitochondria.) We use the allele that has highest frequency in the samples as the reference allele for each position.

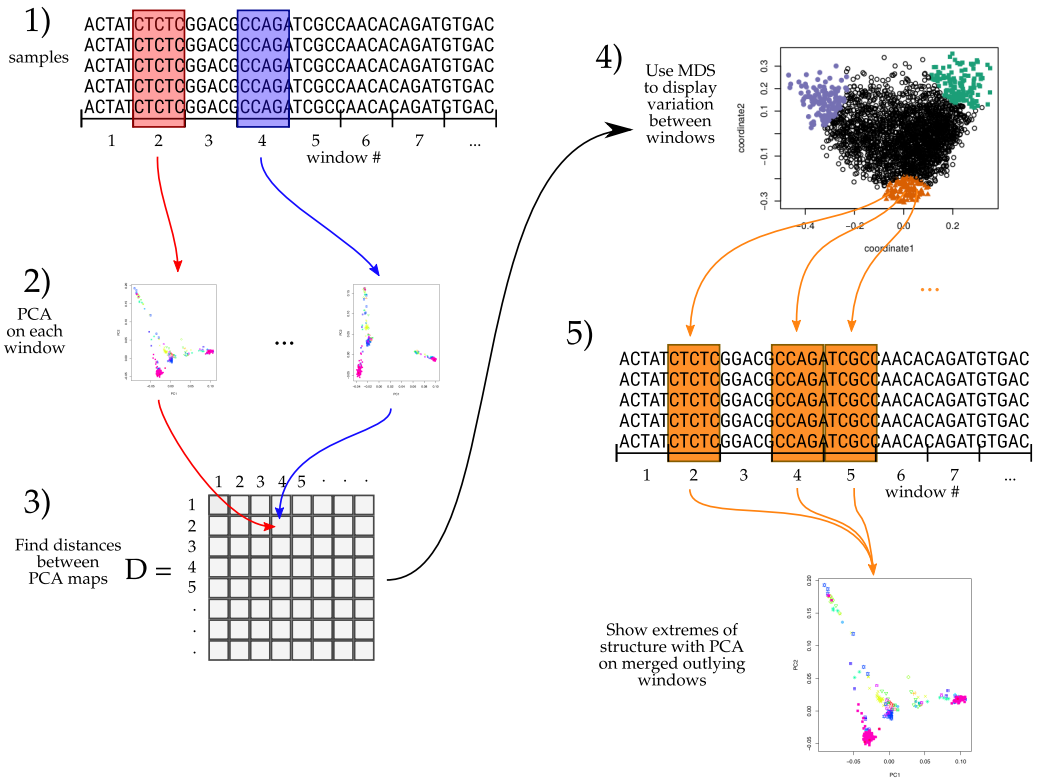


Figure 1: The diagram of our method.

***Drosophila melanogaster*:** We also used whole-genome sequencing data from Drosophila Population Genomics Project (DPGP) and John Pool's lab (Lack et al. 2015), which together has 380 samples from 16 countries across Africa and Europe. *Check: are all of our samples described in Lack et al. (2015)? If yes, just say DPGP; if no, add appropriate citation (ask John what this should be). Also, add website?* Since the *Drosophila* samples are from inbred lines, we treat the samples as haploid when recoding; regions with residual heterozygosity were marked as missing in the original dataset. Each chromosome arm we investigated (Chr2L, Chr2R, Chr3L, and Chr3R) has 2–3 million SNPs. Due to high density of missing data for some parts in the genome, we first deleted any samples with more than 8% missing genotypes, and then deleted positions with more than 20% missing data. (We chose these cutoffs as the tails of the relevant empirical distributions.)

***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; 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
Human	100	636,494	203	0.55
Drosophila melanogaster	1,000	9,019	2,674	0.53
Medicago truncatula	10,000	102,580	467	0.50

Table 1: Descriptive statistics for each dataset used, with the chosen window sizes. See text for how window sizes were chosen; in each window we performed PCA, computed the percent varince explained by the top two principal components; the mean across windows is given in the last column.

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. Specifically, we did PCA as follows: 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. *Is that correct? This is not $X^TX/(m-1)$, which is what was here before, but is what cov(), use='pairwise' does: see* `package/tests/testthat/test_covariance.R`. 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.

I'm talking about flipping signs up here, even though in the code we do it when computing variances, since it makes sense conceptually here. Since eigenvectors are still only defined up to sign, when comparing between windows we choose the sign to best match each other: after choosing $PC1_1$, for instance, and u is the first eigenvector obtained from the covariance matrix for window j, then $PC1_j = \pm u$, where the sign is chosen according to

which of $\sum_i (PC1_{i1} - u_i)$ or $\sum_i (PC1_{i1} + u_i)$ is smaller.

2.3 Choosing window length

The choice of window length entails a balance between several factors. 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. Window length choice therefore entails a signal versus noise tradeoff in the estimates of population structure. 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 $\overline{PC1}_j = (1/N)\sum_{j=1}^{N} PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jackknife (Efron and Efron 1982): we divide the j^{th} block 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 $\sigma^2_{ij} = \frac{9}{10}\sum_{k=1}^{10}(PC1_{ij,k} - \frac{1}{10}\sum_{\ell=1}^{10} PC1_{ij,\ell})^2$, and we measure

noise by averaging over samples and windows:

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

These values are shown calculated separately for each chromosome arm in the *Drosophila* dataset in Table 2. Finally, we choose 100 SNPs, 1000 SNPs and 10000 SNPs as window length for human, *Drosophila*, and *Medicago* respectively.

chrom. arm	window length (in SNPs)	100	500	1,000	10,000	100,000
2L	σ_{noise}	4.53	4.05	3.44	1.30	0.63
	σ_{signal}	5.25	5.19	4.72	2.60	1.77
2R	σ_{noise}	4.67	4.38	4.04	2.40	1.16
	σ_{signal}	5.27	5.20	5.15	4.81	4.27
3L	σ_{noise}	4.56	4.47	4.05	2.71	1.57
	σ_{signal}	5.10	5.02	4.90	4.10	4.35
3R	σ_{noise}	4.42	4.20	3.79	2.42	1.42
	σ_{signal}	5.08	5.01	4.94	4.43	3.74
X	σ_{noise}	4.98	4.52	3.92	4.02	1.30
	σ_{signal}	5.11	4.93	4.80	1.80	3.38

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 100 (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 compared population structure in different genomic windows using the first two principal components (PCs) and the corresponding eigenvalues from PCA. We do this, rather than using the entire covariance matrix for computational efficiency and because the top PCs summarize important population structure, using only these should reduce the effect of noise. For example, the constructed matrixes for ith and jth window are as following.

(λ_{1i} and λ_{2i} are the eigenvalues for the first two PCs for ith window; M_i is the constructed new matrix for ith window and jth window. *Say what M_i and M_j are in the text. Also, how about we write V instead of PC? I like to use only one letter for variables, so it's clear it's just one thing, not the product of P and C. And, isn't it $\sqrt{\lambda_{1j}^2 + \lambda_{2j}^2}$ on the bottom? Check in the R code (pc_dist.R).*

$$M_i = \frac{\lambda_{1i} PC1_i PC1_i^T + \lambda_{2i} PC2_i PC2_i^T}{\lambda_{1i} + \lambda_{2i}} \tag{1}$$

The Euclidean distance D_{ij} between the matrices M_i and M_j stands for the similarity of population structure for the ith window and jth window. Due to the orthogonality of eigenvectors,we could use the following method to calculate the pairwise distance greatly saving time and space.

Define D.

$$V_{i1} = \sqrt{\frac{\lambda_{1i}}{\lambda_{1i} + \lambda_{2i}}} PC1_i \qquad V_{i2} = \sqrt{\frac{\lambda_{2i}}{\lambda_{1i} + \lambda_{2i}}} PC2_i \tag{2}$$

$$M_i = V_{i1} V_{i1}^T + V_{i2} V_{i2}^T \tag{3}$$

$$\begin{aligned} D_{ij} = \Big\{ &(V_{i1} \cdot V_{i1})^2 + (V_{i2} \cdot V_{i2})^2 + (V_{j1} \cdot V_{j1})^2 + (V_{j2} \cdot V_{j2})^2 \\ &- 2 \left[(V_{i1} \cdot V_{j1})^2 + (V_{i1} \cdot V_{j2})^2 + (V_{i2} \cdot V_{j1})^2 + (V_{i2} \cdot V_{j2})^2 \right] \Big\}^{1/2} \end{aligned} \tag{4}$$

Using this procedure, we get the pairwise distance matrix that says how similar popu-

lation structure is in each pair of genomic windows.

2.4.1 Visualize the pairwise distance matrix

We use Multidimensional scaling (MDS) method to visualize the distance matrices. It can reduce the dimensionality of a distance matrix while preserve the distance information between objects as well as possible. The result is a set of coordinates for each sample with the property that the first M coordinates give the arrangement in M-dimensional space that best recapitulates the original distance matrix. We use M=2 to produce one or two dimensional visualization of relationships between windows' population structure.

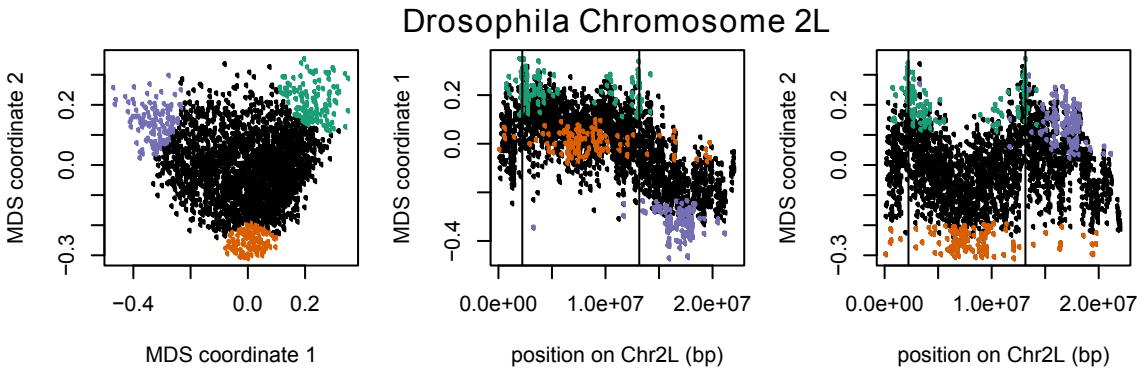
3 Results

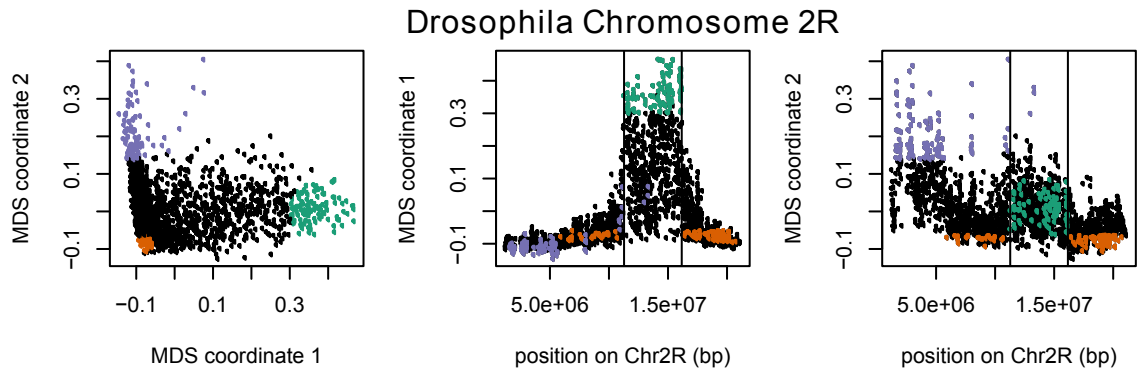
In all these 3 species, PCA plots vary along the genome in a systematic way, showing strong chromosome-scale correlations. This implies that variation is truly due to population structure, since fluctuations due to demographic noise are not expected to show long distance correlations. Below, we display the results and investigate likely underlying causes.

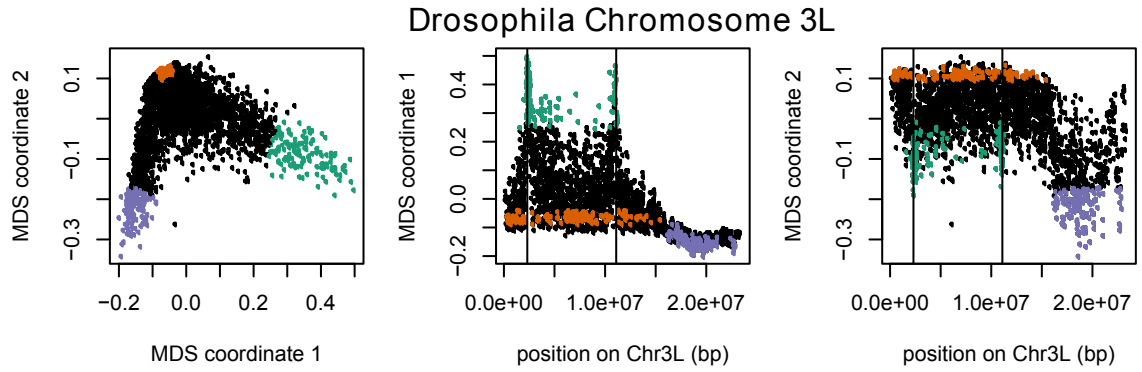
3.1 *Drosophila melanogaster*

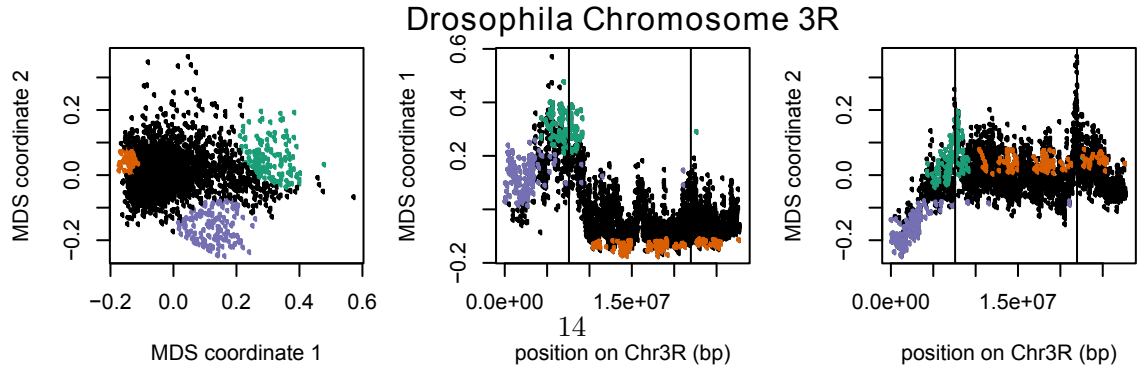
We ran the above method on chromosome arms 2L, 2R, 3L, 3R and X separately. For each, the two-dimensional MDS visualization resembles a triangle. (eg. Figure 2a) Since the relative position of each window in this plot shows the similarity between windows, this suggests that there are 3 extreme types of population structure shown in the 3 peaks of the "triangle", and that other window's population structure might be a mixture of those extremes.

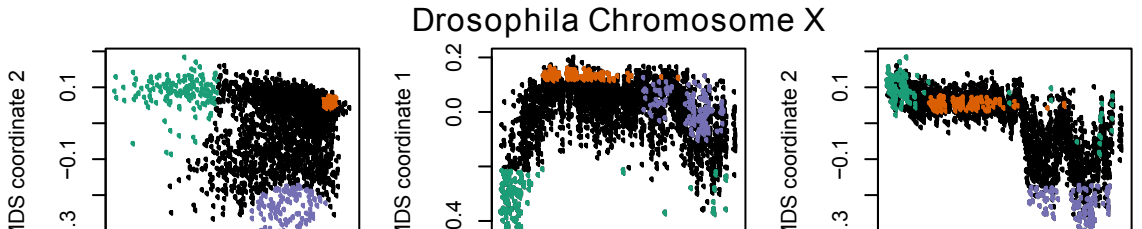
To investigate these extremes, we pick a window for each extreme, and take out the 5% of windows that are closest to it in the MDS coordinates, then combine those windows


Drosophila Chromosome 2L
MDS coordinate 2
MDS coordinate 1
MDS coordinate 1
position on Chr2L (bp)
MDS coordinate 2
position on Chr2L (bp)


Drosophila Chromosome 2R
MDS coordinate 2
MDS coordinate 1
MDS coordinate 1
position on Chr2R (bp)
MDS coordinate 2
position on Chr2R (bp)


Drosophila Chromosome 3L
MDS coordinate 2
MDS coordinate 1
MDS coordinate 1
position on Chr3L (bp)
MDS coordinate 2
position on Chr3L (bp)


Drosophila Chromosome 3R
MDS coordinate 2
MDS coordinate 1
MDS coordinate 1
position on Chr3R (bp)
MDS coordinate 2
position on Chr3R (bp)

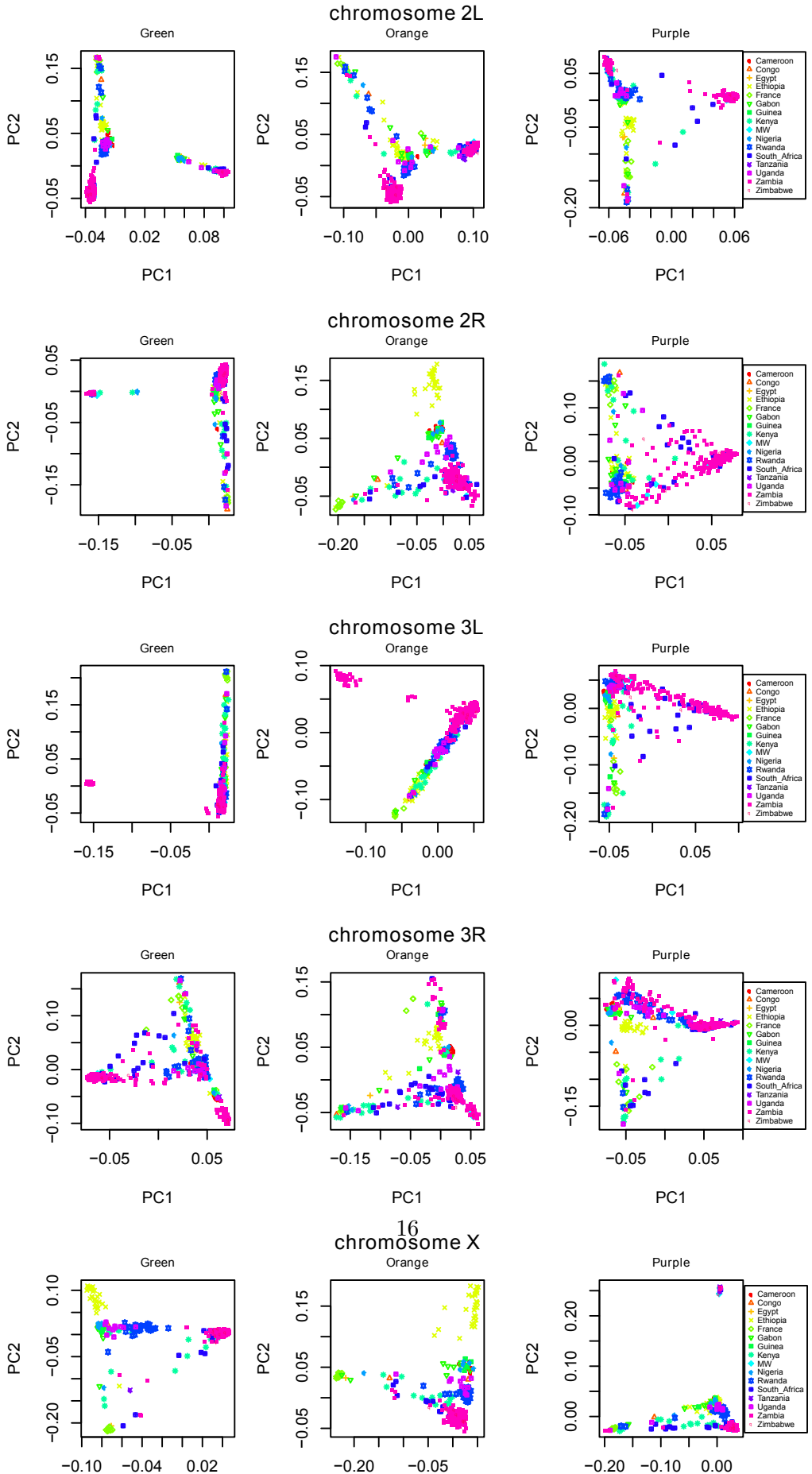

Drosophila Chromosome X
MDS coordinate 2
MDS coordinate 1
MDS coordinate 2

for each extreme and apply PCA on the corresponding sections of genome. We can see in Figure 3 the obvious difference between their PCA plots.

There's a large inversion on Chr2L in that is polymorphic in these samples, In(2L)t (Corbett-Detig and Hartl 2012). We recolored the PCA plots in Figure 3 by the orientation of the inversion for each sample. Figure 4 Two regions of similar, extreme population structure (green in Figure 2) are found around inversion breakpoints, and the other two extremes occur in the center of the inversion and between the inversion and centromere. The corresponding PCA plots show that locally, population structure is mostly determined by which orientation of the inversion each sample has. Similar results are found in other chromosome arms that have known polymorphic inversions (Chr2R, Chr3L, Chr3R) in *Drosophila*. Other known polymorphic inversions are like In(3L)P on Chr3L, In(3R)Mo and In(3R)P on Chr3R. In(3L)P and In(3R)Mo are just a few in our samples, thus we don't see the significant influence of population structure of them. The situation of Chr3R is a little complicated. The coexisting inversions make the MDS visualiazation of it more difficult to pick the extremes. For the extremes we picked in Figure 2, In(3R)K could best explain the PCA clustering result, while picking other windows as extremes might show that In(3R)P explain the PCA clustering result.

3.2 Human

We ran our method separately on all 22 human autosomes. For instance, the eleven windows that are outliers in the first MDS coordinate of chromosome 8 (Figure 5b) coincide with the position of known polymorphic inversions on 8p23. Similar results are found in other chromosomes that have known inversions (eg. Chromosome 15, Chromosome 17). We found that the primary axis of variation in population structure differentiated only one or a few windows on each chromosome, unlike the countinous variation in population


chromosome 2L
Green
Orange
Purple
chromosome 2R
Green
Orange
Purple
chromosome 3L
Green
Orange
Purple
chromosome 3R
Green
Orange
Purple
chromosome X
Green
Orange
Purple
PC1
PC2
Cameroon
Congo
Egypt
Ethiopia
France
Gabon
Guinea
Kenya
MW
Nigeria
Rwanda
South_Africa
Tanzania
Uganda
Zambia
Zimbabwe

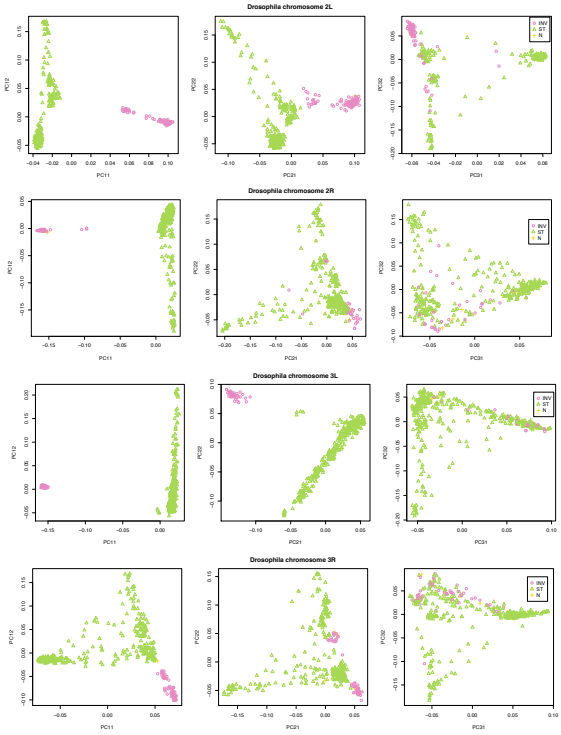


Figure 4: As in Figure 2, except that samples are colored by orientation of the corresponding polymorphic inversion, In(2L)t, In(2R)NS, In(3L)OK and In(3R)K. (Data from Lack et al. (2015).

structure we see in Drosophila. Other chromosomes showed similar results around predicted inversions: PCA might provide an additional way to identify inversions (Ma and Amos 2012).

When we run the method on all 22 autosomes together, the outlying signal of chromosome 8 is still visible. (See supplementary for all 22 autosomes)

3.3 *Medicago truncatula*

We ran our method on all 8 chromosomes of *Medicago truncatula* seperately and found the correlation between each MDS plot and gene density along the genome in all 8 chromosomes. This consistency implies that the factor driving the population structure for each chromosome might be the same, for example, background selection. So we ran all chromosomes together, that is, calculating the pairwise distance for all the windows along 8 chromosomes, and then get the MDS coordinates for each chromosome by locating the whole MDS coordinates. *The results looked different than the other two species, with much less pronounced peaks...* We computed gene density near each window using gene models in Mt4.0 JBrowse (Tang et al. 2014). The first MDS coordinate value is negatively correlated to the gene count for each window (Figure 9). We found the position of the peak for each MDS plot has a coincidence with the position of heterochromatic regions. This means that population structure in the windows located in heterochromatin tends to have higher similarity, since those windows are closer in MDS plots. Biologically, heterochromatic regions have lower gene density and may be less subject to selection (Kulikova et al. 2001; Paape et al. 2013). Besides, we found the PCA plots are quite consistent for the peaks that more clustering away from the heterochromatic regions, while the PCA plots are barely consistent for the peaks that more clustering around heterochromatic regions for the 8 chromosomes. Unlike in the human and Drosophila genomes, in *Medicago truncatula*,

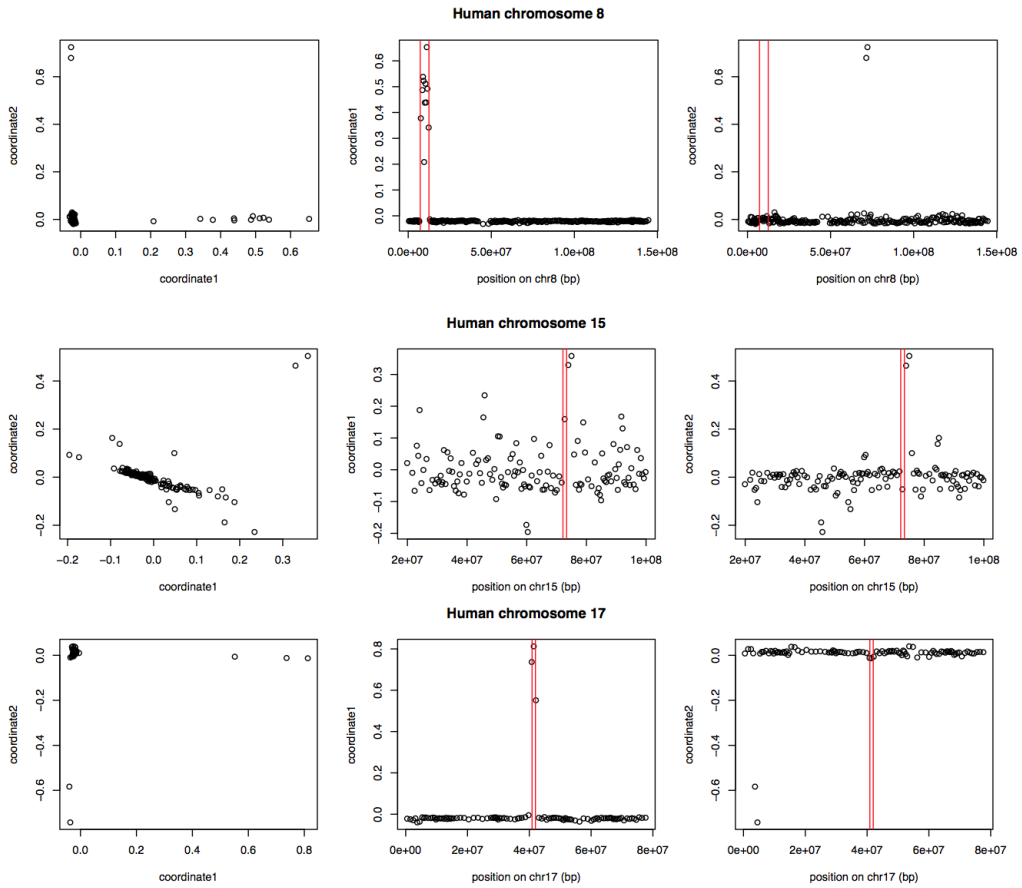


Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in the plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position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

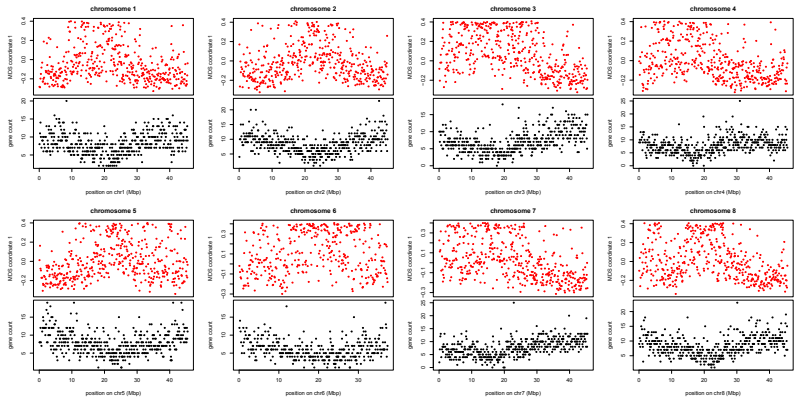


Figure 6: MDS results and gene density for the *Medicago* genome (chromosome 1-8). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene count for each window against the position of each window.

major variation in population structure is likely due to linked selection.

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 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. It is unclear whether the technique

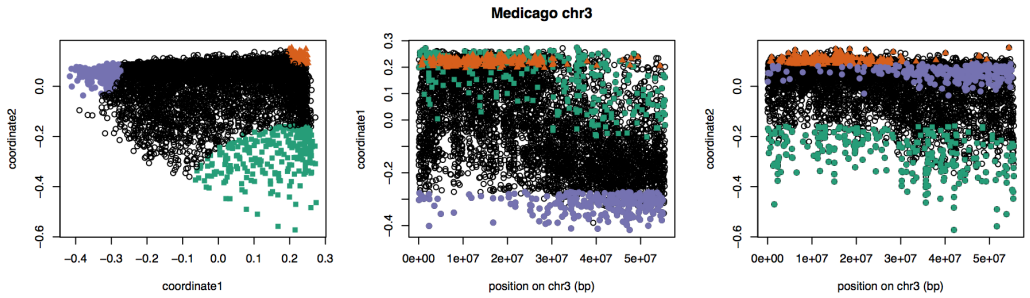


Figure 7: MDS visualization for Medicago chromosome 3. Each point in the plot stands for a window (length 1000 SNPs).

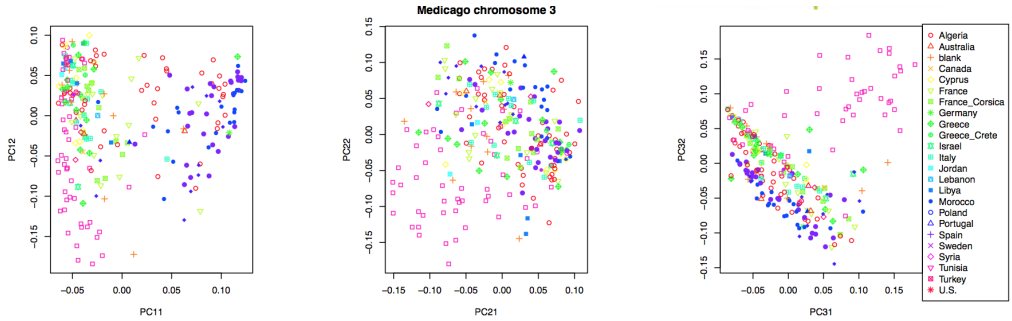


Figure 8: PCA plots for the three sets of genomic windows colored separately in Figure 7. From left to right, the windows colored green, orange, and purple. Each point corresponds to a sample, colored by origin.

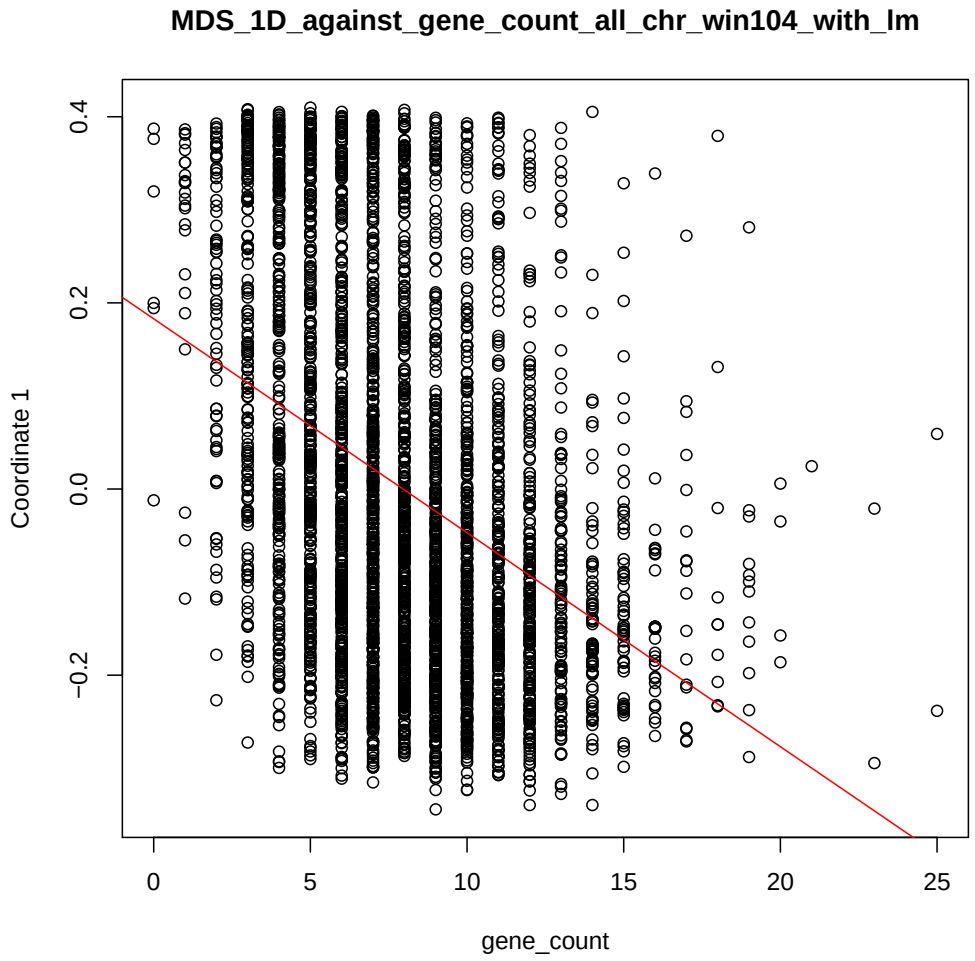


Figure 9: First MDS coordinate against gene density for all 8 chromosomes. The first MDS coordinate is significantly correlated with gene count (r=0.146, p=2.2×10^{-16}). (See supplementary for each single chromosome's MDS result against gene count for *Medicago*)

will be useful on reduced representation genotyping datasets due to marker density and issues with missing data – our investigations with one such dataset were inconclusive – but even low coverage, whole-genome sequence is very promising.

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 (**??**). Since population structure is not constant along the genome, this could in principle lead to an inflation of false positives 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.

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. Next, the choice of window size may be important, although in our applications results were not sensitive to this, indicating that the limit of resolution was smaller than the scale on which patterns of kinship varies along the genome. 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 amount of phylogenetic information. However, given the insensitivity of our results to window length, this seems unlikely to give different results. *If we check: We did not have this choice to have a substantial effect on results.*

More generally, there are many possible methods to discover common structure in different parts of the genome. The methods we chose discovered strong biological signal of different types in three datasets; but it is 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. Minor points we have not explored include how to decide how many PCs to use in approximating structure of each window (equations XX), how many MDS coordinates to use when describing the distance matrix between windows, or how to choose interesting regions of the genome when the MDS plot is not triangular. 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. 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 smoehow removing the effects of inversions in the *Drosophila melanogaster* or human datasets would produce a pattern similar to that seen in *Medicago truncatula*.

Check if I've missed anything, then remove these?

4.1 Future work

1. For human and Drosophila, we want to eliminate the regions under known inversions and check the variation of population structure for the remaining part by removing those sections. We try to check whether they will give similar results as in *Medicago truncatula*, that is whether the variation is closely related to heterochromatin or gene density.

2. Uneven sampling has a strong influence on PCA projections (McVean 2009). Our human data, POPRES, is unevenly sampled including 346 African-Americas, 73 Asians, 359 Indian Asians and 3187 Europeans. First, we'll try sub-sampling Europeans to balance the population size for the 4 population and repeat the process on the resampled data. Second, we'll try to apply the whole process on only European samples to see the genetic variation inside European samples. Third, we want to try different scheme of adding a weighting matrix to the covariance matrix of genotype data, thus to the reduce the influence of uneven sampling.

3. Since regions that have low recombination rate tend to have similar PCs, we'll try cutting the chromosomes into windows with same distance in genetic map instead of same SNP numbers.

4. Euclidean distance between the contracted matrix based on PCs is one measure of the similarity for window's population structure. We want to try other methods of distance between windows, for example, we used the distance for PCs to reduce noise, however the distance between covariance matrixes of genotype matrix might also be informative.

5. Although the first two coordinates contains the main part of information, we'd like to see the information contained in higher PCs (e.g. the third PC, the forth PC), and higher dimension of MDS.

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`.

N 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/`.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Gideon Bradburd, Peter L. Ralph, and Graham Coop. A spatial framework for understanding population structure and admixture. *bioRxiv*, January 2015. URL `http://biorxiv.org/content/early/2015/01/07/013474`.

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`.

B 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.

Bradley Efron and B Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

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): e1001117, 2010.

D Falush, M Stephens, and J 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?dopt=Abstract.

Ziyue Gao, Molly Przeworski, and Guy Sella. Footprints of ancient balanced polymorphisms in genetic variation data, 2014. URL http://arxiv.org/abs/1401.7589. cite arxiv:1401.7589.

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.

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, 2013. URL http://arxiv.org/abs/1302.1148. cite arxiv: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. Natural selection shapes the mosaic ancestry of the Drosophila Genetic

Reference Panel and the *D. melanogaster* reference genome. *bioRxiv*, 2015. doi: 10.1101/014837.

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.

Sewall Wright. The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354, 1949. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1949.tb02451.x. URL `http://dx.doi.org/10.1111/j.1469-1809.1949.tb02451.x`.

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`.

.1 Weighted PCA

PCA 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 $N \times N$ matrix of genotypes, 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

$$\|C - U\Lambda U^T\|^2 = \sum_{ij} \left(G_{ij} - \sum_m \lambda_m U_{im} U_{jm} \right)^2. \tag{5}$$

The columns of U, known as the principal components, are the eigenvectors of C, the eigenvalues of C are λ, 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 (5) Let w_i be a weight associated with sample i, W the diagonal matrix with w along the diagonal, and

instead seek to minimize

$$\|W^{1/2}(C - U\Lambda U^T)W^{1/2}\|^2 = \sum_{ij} W_i W_j \left(G_{ij} - \sum_m \lambda_m U_{im} U_{jm} \right)^2, \tag{6}$$

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 (6), let λ and V denote the eigenvalues and eigenvectors of $W^{1/2}CW^{1/2}$, so that $V\Lambda V^T$ is the closest in least squares to $W^{1/2}CW^{1/2}$; so if we define $U = W^{-1/2}V$ then $U^T W U = V^T V = I$, and

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

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

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

Note finally that since we only need find the top k eigenvectors; in our R implementation we use the Spectra library (Qiu and Mei 2016).