Local PCA Shows How Population Structure Differs Along the Genome

Han Li, Peter Ralph

May 25, 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, correlates with local gene density, and may be caused by background selection or local adaptation.

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 kinship matrix contains the kinship coefficient for pairwise individuals. Kinship coefficient defines the genetic relatedness between individuals. It is the probability that two alleles randomly selected from two individuals are inherited from the most recent common ancester. It could be estimated from given pedigree or from genome-wide covariances of genotype markers. It is well-known that for kinship matrix, actual relatednesses have a lot of noise about the expected value, and depend on where on the genome you look; this is why scans for selective sweeps work. Populations are often structured in some way while there are systematic genetic variation between populations.

About 37 years ago, Menozzi et al. (1978) first applied principal component analysis (PCA) in population genetics to construct maps summarizing genetic variation (Menozzi et al. 1978). Nowadays, PCA is a widely used powerful non-parametric method to extract information from genetic data. PCA results are derived from the covariance matrix of genotype matrix. The results of PCA can be directly related to the underlying genealogical history of the samples, such as coalescence time (time to most recent common ancestor) and migration rate between populations (McVean 2009; Novembre and Stephens 2008). Through dimension-reduction, PCA can identify key components of population structure, which describes how different samples are related, and are often closely related to geography. Plots of the first two principal components (PCs) can mimic the samples' geographic origin to some extent. Since population structure describes how different sam-

ples are related, samples living closer tend to be more genetically similar and thus tend to be clustered in PC plots (Novembre et al. 2008; Patterson et al. 2006). However this relatedness is limited while there's recent migration or for group with nongeographic kinship patterns,for example, social or religious groups. (Astle and Balding 2009)

PCA is often used in genome-wide association studies (GWAS) for stratification correction (Price et al. 2006). However, different parts of genome have different genetic features. First, each site of DNA may have different gene tree. The covariance matrix of genotype matrix averages those gene trees. For a genomic region, if individuals have alleles closer in gene tree, they tends to be more close in PC projections for that region. Different DNA segments may have different gene tree and therefore different population structure for those segments. Second, the strength of linked selection differs for different DNA segments, and produces different population structure in region under linked selection compared to other region. Selective sweeps cause local recent ancestry or short trees. Balancing selection causes deep trees. Background selection causes shallow ones. Third, if a chromosome inversion is polymorphic in the sample, the regions around the breakpoints of inversions usually have high linkage disequilibrium and the two directions of a inversion will have different linked alleles around the breakpoints. Recombination suppression across inversions thus results in different genome structure and population structure. Other effects, like noise, introgression might also influence population structure.

Investigating the genomic variation along the genome can help us to have a better understanding of the relation between genome structure and population structure, and could possibly lead to more powerful methods for GWAS.

To investigate this, we cut each chromosome into windows (with hundreds to thousands of SNPs in each), applied PCA to each window, and visualized how population structure, as summarized by PCA, varies along windows.

In this project, we used SNP data for human, *Medicago truncatula*, and whole genome sequencing data for *Drosophila*. Based on the principal components, we can estimate the similarity of population structure contained in each genome window. To quantify similarity of population structure between windows, we constructed for each window an approximate, scaled covariance matrix based on the first two PCs measured the pairwise Euclidean distance between those matrices. We 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). To interpret the results of MDS, we combine known genome feature information for each species, such as the distribution of inversions, and heterochromatin and gene density along the genome. Each species showed distinct patterns, reflecting differences in their biology.

Improve this paragraph: Other methods for visualizing population structure are like STRUCTURE, (Falush et al. 2003, 2007; Hubisz et al. 2009; Pritchard et al. 2000) model-based approach, (Yang et al. 2012) maps of heterozygosity (Ramachandran et al. 2005).

A number of 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 Other Introduction

Catchy start: the kinship matrix goes back to almost Mendel and is essential in GWAS; however, it is well-known that actual relatednesses have a lot of noise about the expected value, and depend on where on the genome you look; this is why scans for selective sweeps work.

Review of kinship matrix: it's either an expected kinship, given the pedigree; or an estimated genome-wide average. Wright's path coefficients (Wright 1943). Why it helps with confounding for GWAS. Graham & Vince's paper, maybe. Like IBD, is only well-defined in a known pedigree, up to the founders. Kinship and confounding reviewed in Astle and Balding (2009).

Kinship matrices differ for sex chromosomes and the like.

Review of selection causing differential patterns along the genome. Locally everything is treelike (gene trees); kinship matrix is an average of these (write equation for this). Selective sweeps cause local recent ancestry/short trees. Balancing selection causes deep trees. Background selection, shallow ones. Extreme examples of free gene flow in some places between species: e.g. Heliconius. Introgression may be nonuniform, e.g. neanderthal, others. Refs: hitchhiking (Maynard Smith and Haigh 1974), Kim and Maruki (2011) study hitchhiking in a spatially subdivided population, McVean (2007) looks at the effect of selection on LD. Barton (2000) reviews hitchhiking. Bierne (2010) discusses how hitchhiking effect decreases with geographic distance. Charlesworth et al. (2003) reviews patterns of diversity, relating spatial structure to effects of selection.

Review of methods looking along the genome: argweaver, HMM between species, ???

What is "population structure"? Asks which "populations" are closely related, more diverged, how much diversity do they harbor. Often geographic. Vital in exploratory data analysis. It is a summary of kinship: lack of migration between pops causes a deficit in

connections through the pedigree, and so affects kinship. Wright defined F_{ST} in (Wright 1949), and says "It has probably occurred to the reader that the coefficientof inbreeding may mean very different things in different cases."

Review of methods for visualizing pop structure: PCA, structure (Falush et al. 2003), EEMS, (Petkova et al. 2014), (Yang et al. 2012), Maps of heterozygosity (Ramachandran et al. 2005). Genealogical interpretation of PCA by McVean (2009). Other semi-related stuff: estimation of covariance matrices; local pca(?);

3 Methods

How about: first describe the method; then afterwards, present the three datasets.

3.1 Recode the DNA sequence to a matrix consisting of 0,1,2 (and NA).

For human, we use SNP chip data from POPRES (Nelson et al. 2008). There are 3,965 samples in total, (346 African-Americas; 73 Asians; 3,187 Europeans; 359 Indian Asians), and the 22 autosomes together have 447,267 SNPs in this dataset. We use the allele that has highest frequency in the samples as the reference allele for each position. The first step is to recode genetic data as a numeric matrix; which we take to have one row per variant and one column per sample. (omitting monomorphic variants) In our applications, for each variant we picked a reference allele, and recoded genotypes as the number of non-reference alleles carried by the individual; obtaining 0,1, or 2 for autosomal variants in diploids, or NA for missing data. A normalization step (see below) ensures the result does not depend on the choice of reference allele.

For *Drosophila melanogaster*, we use the 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. Each chromosome arm we investigated

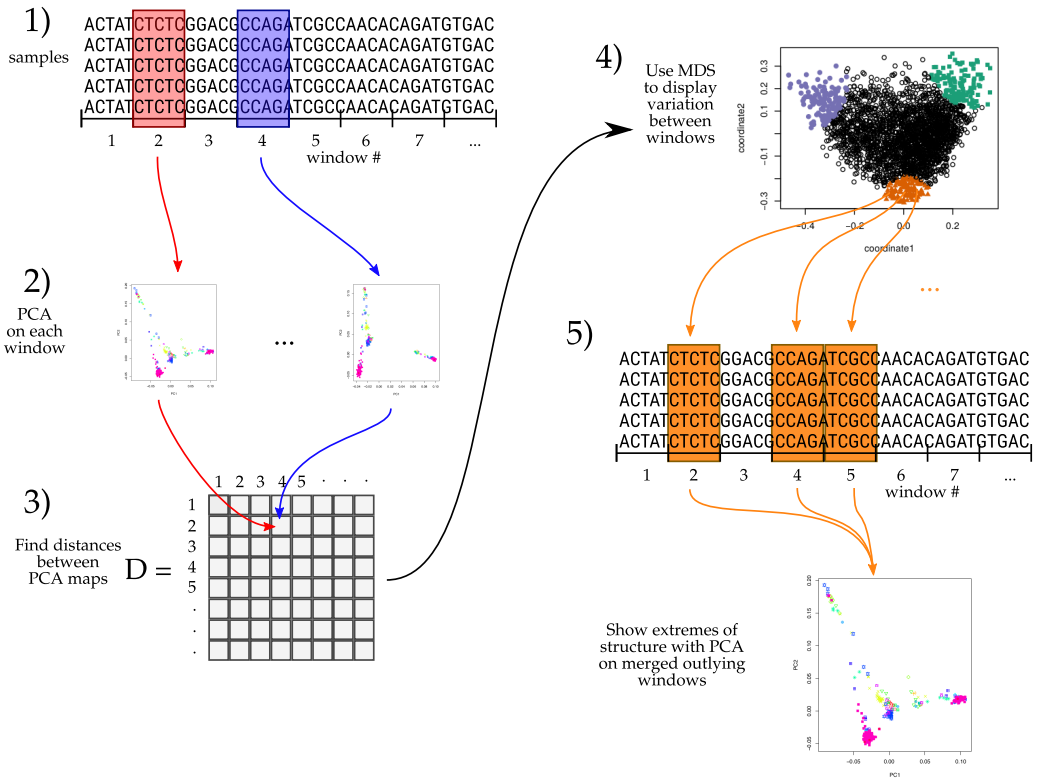


Figure 1: The diagram of our method.

(Chr2L, Chr2R, Chr3L, Chr3R) has 2-3 million SNPs. Due to high density of missing data for some parts in the genome, we then delete the samples with more than 8% missing genotypes and then deleting positions with more than 20% missing data. The cutoff points 8% and 20% are determined from the corresponding distributions of NAs in samples and at positions. Since the *Drosophila* samples are from inbred lines with little heterozygosity, we treat the samples as haploid when recoding.

For *Medicago truncatula*, we use the SNP data from whole genome sequencing from *Medicago truncatula* Hapmap Project (Tang et al. 2014). It has 263 samples from 24 countries. Each of the 8 chromosomes has 3-5 million SNPs.

	species number of SNPs per window	mean length in bp per window	mean number of windows per chromosome	mean ratio of variance explained by top 2 PCs
Drosophila melanogaster	1000	9019	2674	0.53
Human	100	636494	203	0.55
Medicago truncatula	10000	102580	467	0.50

Table 1: some statistic for each species' genome

3.2 PCA in genomic windows

After recoding, we divided the recoded matrix into contigous matrixes that have the same columns but fewer rows than the original matrix, and applied Principal Component Analysis (PCA) separately on each window, following the procedure in McVean (2009) as follows. Starting with the recoded genotype matrix Z, where Z is a $L \times N$ matrix (L is SNP number ; N is sample size), first compute the mean-centered matrix X, $X_{si} = Z_{si} - \overline{Z_s}$, where $\overline{Z_s}$ is the mean of non-missing entries, $\overline{Z_s} = \frac{1}{n}\sum_{j=1}^{n} Z_{sj}$. The mean-centering step makes the result not depend on the choice of reference allele. Then find the covariance matrix

of X, denoted $C = \frac{1}{n-1}XX^T$. We compute the covariance matrix using the R function cov() , with use="pairwise", which computes the covariance between each pair of individuals using all complete pairs of SNPs on those individuals. The principal components are eigenvector of C, 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).

3.2.1 Choosing window length

The window length should neither be too long or too short. In general, longer windows have more SNPs, the noise (the mean standard error within a window) will be smaller and so population structure in that window will be more accurately estimated. However, to better resolve features along the genome, we need reasonably short windows to have stronger signal (the mean standard variation between windows). If we use the first principal component as a measure of population structure, then to choose a proper length for a window, we need to find a balance between variance of the first principal components inside a window and that between windows. The variance between windows is estimated by the mean variance of the first principal component for each window. $Var_{between} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{K}\sum_{i=1}^{K}\left(PC1_{i,j} - \overline{PC1_{,j}}\right)^2\right)$ (K is the number of windows on the genome) The variance inside a window is estimated using the block jackknife cutting the window into 10 equal size smaller windows (Efron and Efron 1982). $Var_{inside} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{9}{10}\sum_{i=1}^{10}\left(PC1_{i,j} - \overline{PC1_{,j}}\right)^2\right)$ Table 2 shows the comparison of variance within a window and that between windows for chromosome arms in *Drosophila*. Finally, we choose 100 SNPs, 1000 SNPs and 10000 SNPs as window length for human, *Drosophila*, and *Medicago* respectively.

	window length(in SNPs)	100	500	10^3	10^4	10^5
Chr2L	SE^2(within)	2.05e-03	1.64e-03	1.18e-03	1.68e-04	4.02e-05
	Var(between)	2.76e-03	2.69e-03	2.23e-03	6.74e-04	3.12e-04
Chr2R	SE^2(within)	2.18e-03	1.92e-03	1.63e-03	5.76e-04	1.35e-04
	Var(between)	2.78e-03	2.70e-03	2.65e-03	2.31e-03	1.82e-03
Chr3L	SE^2(within)	2.08e-03	2.00e-03	1.64e-03	7.32e-04	2.45e-04
	Var(between)	2.60e-03	2.52e-03	2.40e-03	1.68e-03	1.89e-03
Chr3R	SE^2(within)	1.95e-03	1.76e-03	1.44e-03	5.87e-04	2.03e-04
	Var(between)	2.58e-03	2.51e-03	2.44e-03	1.96e-03	1.40e-03
ChrX	SE^2(within)	2.48e-03	2.04e-03	1.54e-03	1.62e-03	1.68e-04
	Var(between)	2.61e-03	2.43e-03	2.30e-03	3.24e-04	1.14e-03

Table 2: Comparison of variance within a window and that between windows for chromosome arms in *Drosophila*.

3.2.2 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).*

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

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.

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

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

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

Using this procedure, we get the pairwise distance matrix that says how similar population structure is in each pair of genomic windows.

3.2.3 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.

4 Results

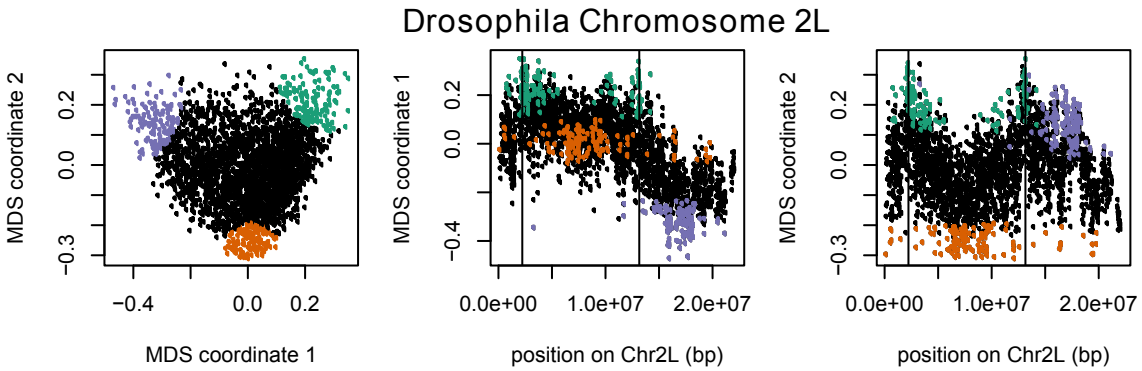
In all these 3 species, PCA plots vary along the genome in a systematic way, showing strong chromosome-scale correlations. This implies that variation is truly due to population structure, since fluctuations due to demographic noise are not expected to show long distance correlations. Below, we display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

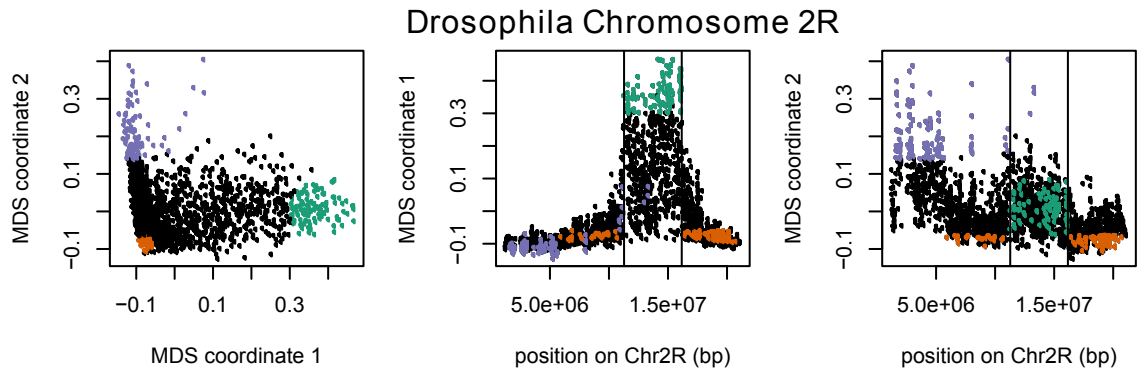
4.1 *Drosophila melanogaster*

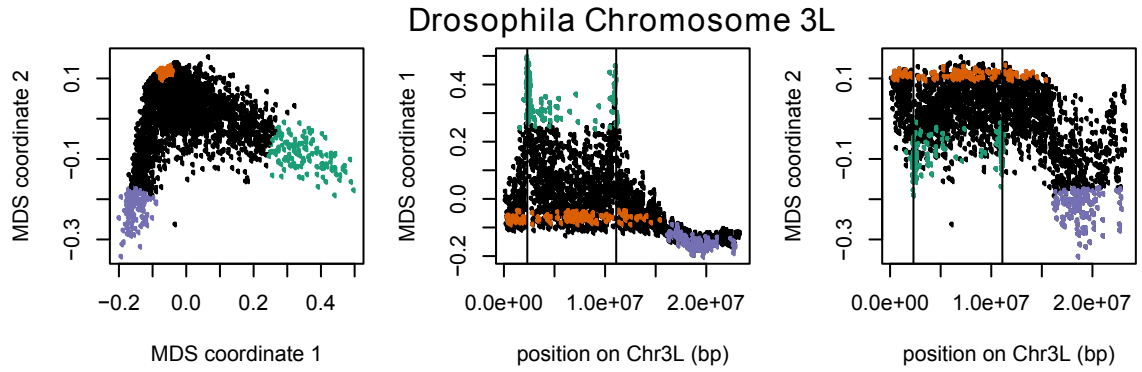
We ran the above method on chromosome arms 2L, 2R, 3L, 3R and X separately. For each, the two-dimensional MDS visualization resembles a triangle. (eg. Figure 2a) Since the relative position of each window in this plot shows the similarity between windows, this suggests that there are 3 extreme types 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.

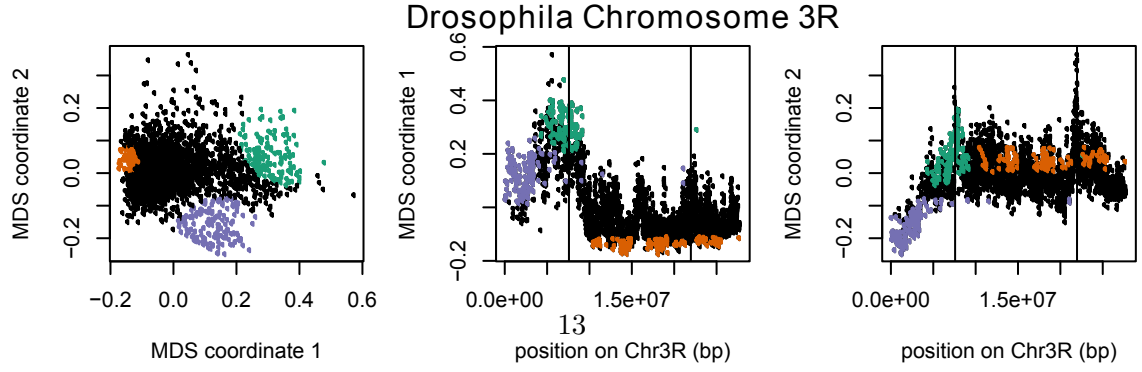
To investigate these extremes, we pick a window for each extreme, and take out the 5% of windows that are closest to it in the MDS coordinates, then combine those windows for each extreme and apply PCA on the corresponding sections of genome. We can see in Figure 3 the obvious difference between their PCA plots.

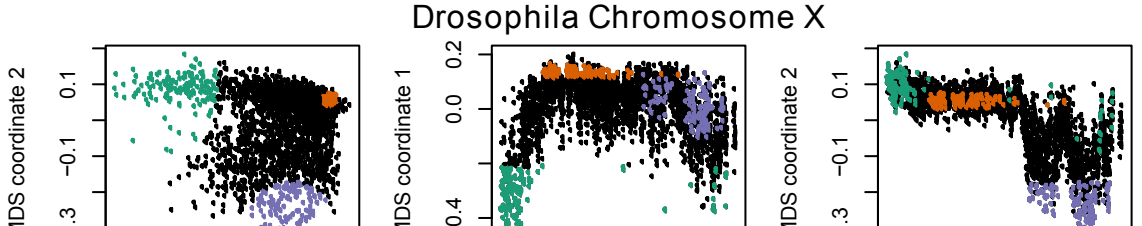
There's a large inversion on Chr2L in that is polymorphic in these samples, In(2L)t (Corbett-Detig and Hartl 2012). We recolored the PCA plots in Figure 3 by the orientation of the inversion for each sample. Figure 4 Two regions of similar, extreme population structure (green in Figure 2) are found around 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

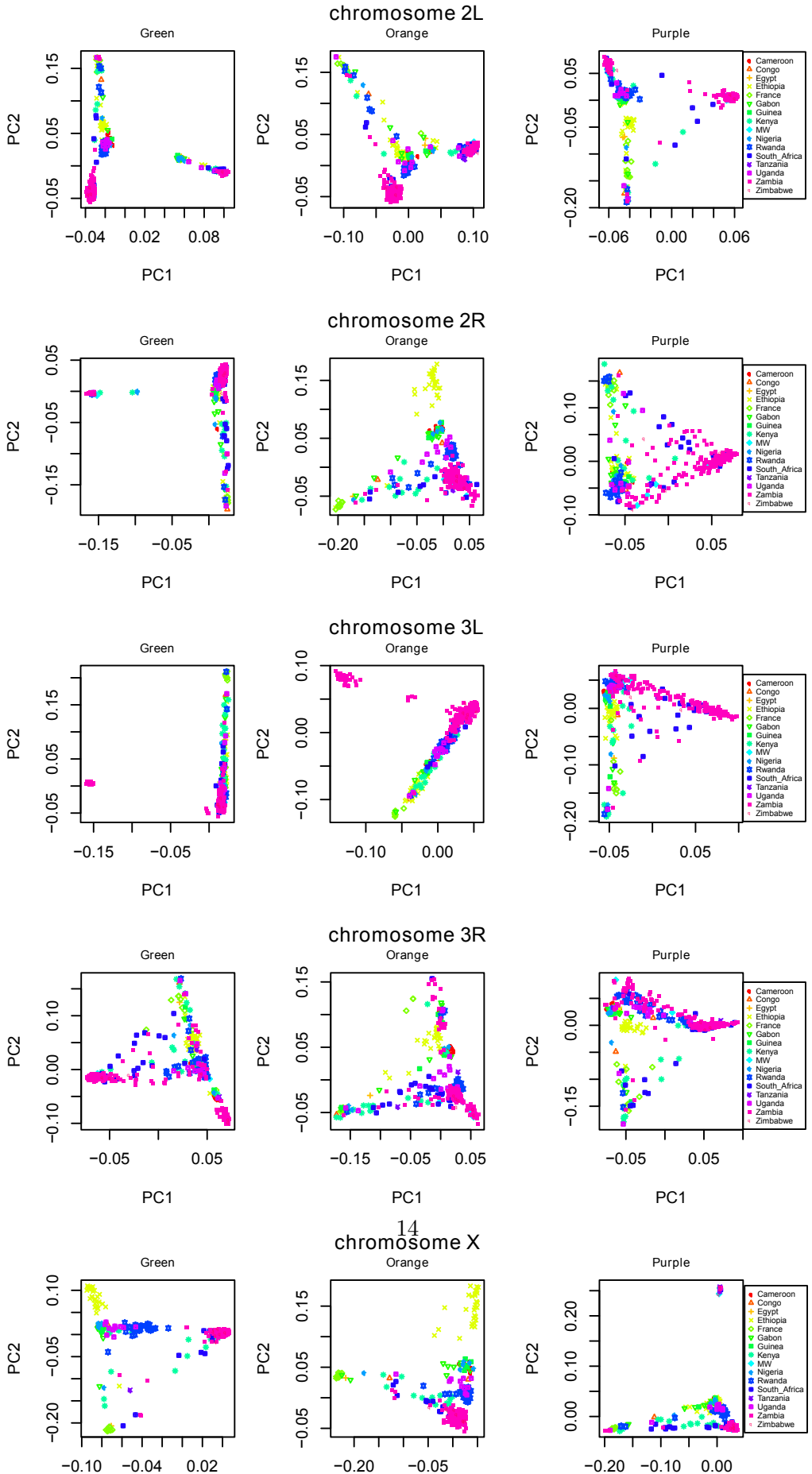

Drosophila Chromosome 2L
MDS coordinate 2
MDS coordinate 1
MDS coordinate 1
position on Chr2L (bp)
MDS coordinate 2
position on Chr2L (bp)


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 1
MDS coordinate 2


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

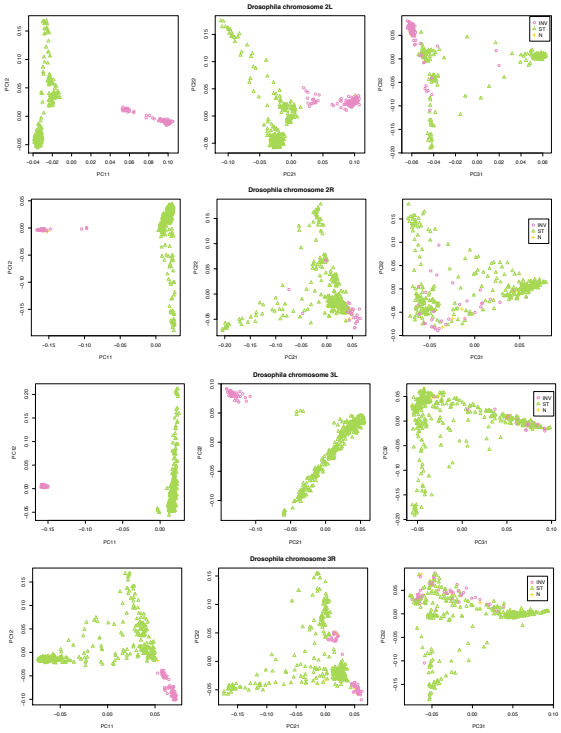


Figure 4: As in Figure 2, except that samples are colored by orientation of the corresponding polymorphic inversion, In(2L)t, In(2R)NS, In(3L)OK and In(3R)K. (Data from Lack et al. (2015).

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.

4.2 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 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)

4.3 *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

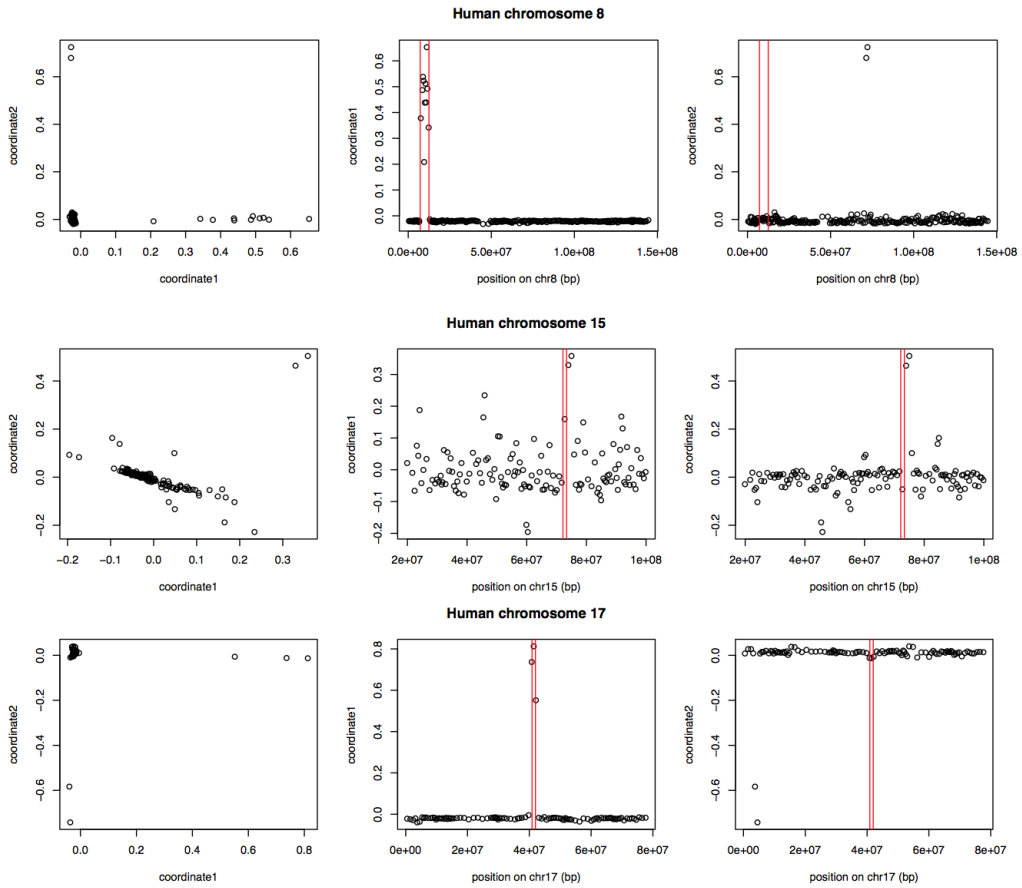


Figure 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in the plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position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

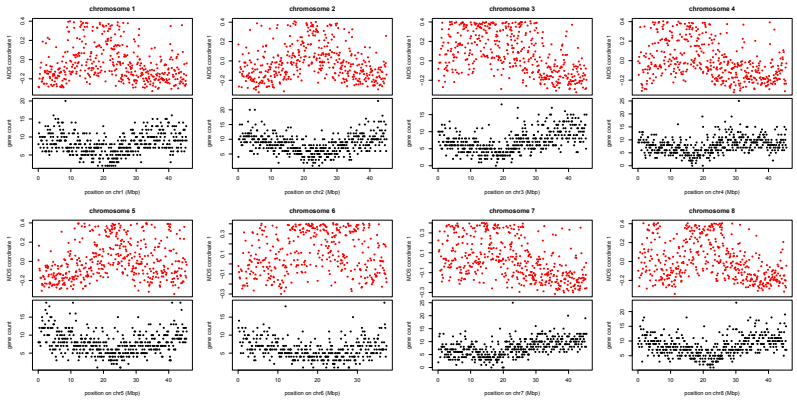


Figure 6: MDS results and gene density for the *Medicago* genome (chromosome 1-8). For each chromosome, the red plot above is first coordinate of MDS against the middle position of each window along each chromosome. The black plot below is gene count for each window against the position of each window.

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

5 Discussion

The phrase "population structure" refers to reduced gene flow between subpopulations, often because of geographical isolation. This leads to systematic patterns in genome-wide mean kinship, and so visualizations of kinship are said to depict population structure,

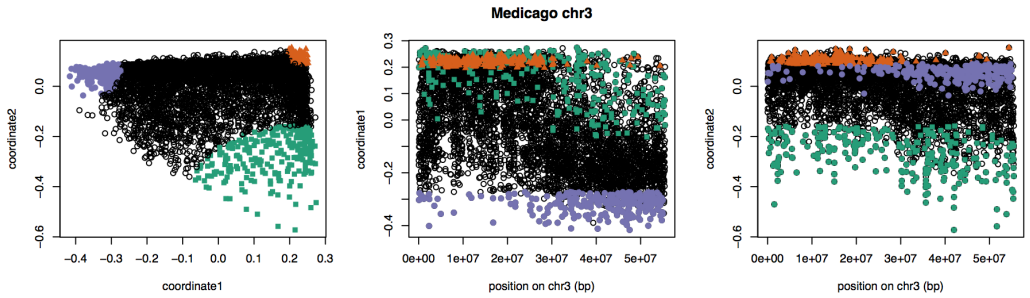


Figure 7: MDS visualization for Medicago chromosome 3. Each point in the plot stands for a window (length 1000 SNPs).

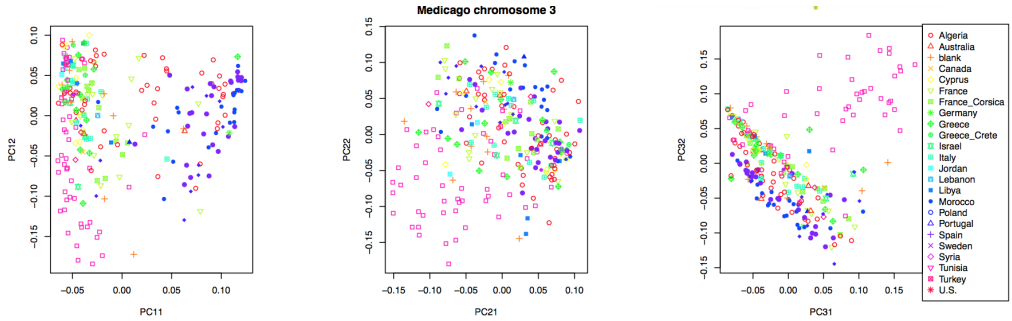


Figure 8: PCA plots for the three sets of genomic windows colored separately in Figure 7. From left to right, the windows colored green, orange, and purple. Each point corresponds to a sample, colored by origin.

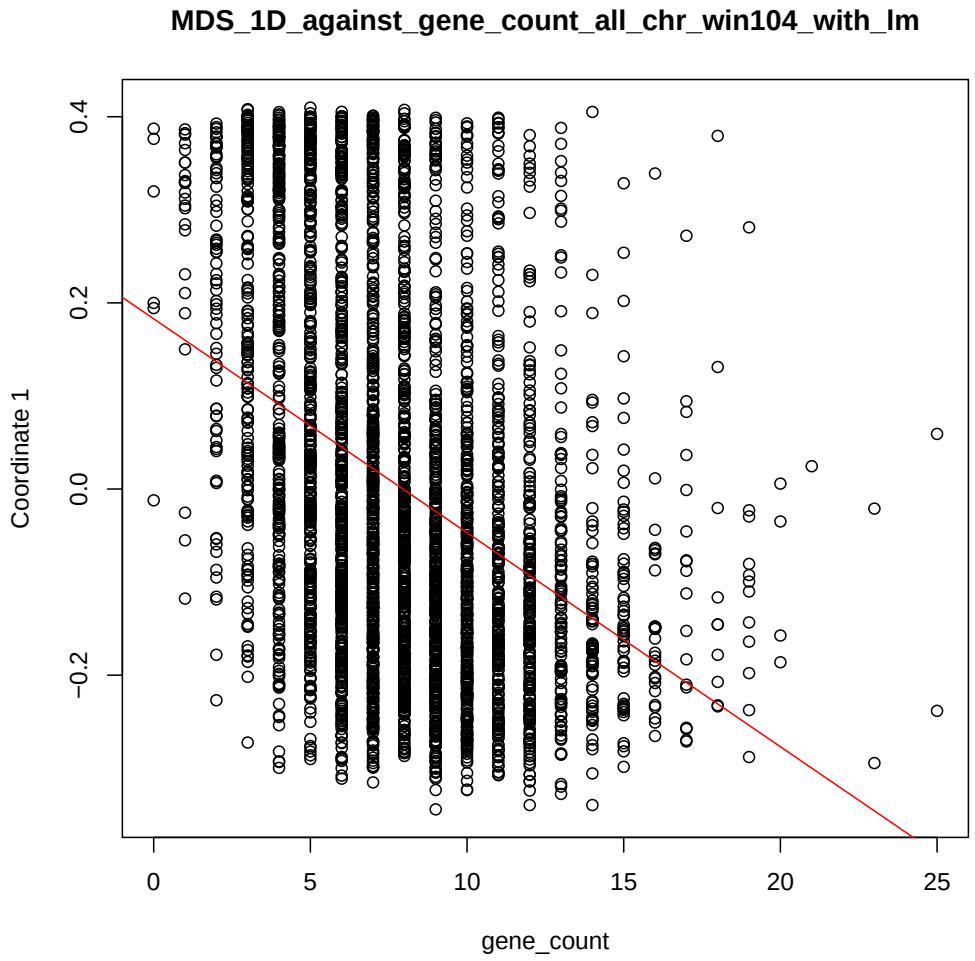


Figure 9: First MDS coordinate against gene density for all 8 chromosomes. The first MDS coordinate is significantly correlated with gene count (r=0.146, p=2.2×10^{-16}). (See supplementary for each single chromosome's MDS result against gene count for *Medicago*)

rather than depicting the *effects* of population structure (which would be more accurate). 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 vary significantly depending on factors including local gene density (**??**). This implies that the population structure of a species depends on which part of the genome is being examined.

With growing appreciation of the heterogeneous effects of selection across the genome, especially the importance of adaptive introgression, hybrid speciation (**???**), local adaptation (**?**), and inversion polymorphisms (**??**), local PCA may prove to be a useful exploratory tool to discover important genomic features. It is unclear whether the technique 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 (**?**); 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 (**?**) 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?

5.1 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/`.

N Bierne. The distinctive footprints of local hitchhiking in a varied environment and global

hitchhiking in a subdivided population. *Evolution*, June 2010. doi: 10.1111/j.1558-5646.2010.01050.x. URL http://www.ncbi.nlm.nih.gov/pubmed/20550573.

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

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.

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.

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/.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes*, 7(4):574–578, 2007.

Melissa J Hubisz, Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources*, 9(5):1322–1332, 2009.

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.

Y Kim and T Maruki. Hitchhiking effect of a beneficial mutation spreading in a subdivided population. *Genetics*, 189(1):213–226, September 2011. doi: 10.1534/genetics.111.130203. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3176130/.

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.

Jianzhong Ma and Christopher I Amos. Investigation of inversion polymorphisms in the human genome using principal components analysis. *PloS one*, 7(7):e40224, 2012.

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.

G McVean. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175 (3):1395–1406, March 2007. doi: 10.1534/genetics.106.062828. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1840056/?tool=pubmed`.

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

Paolo Menozzi, Alberto Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):786–792, 1978.

Matthew R Nelson, Katarzyna Bryc, Karen S King, Amit Indap, Adam R Boyko, John Novembre, Linda P Briley, Yuka Maruyama, Dawn M Waterworth, Gérard Waeber, et al. The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3): 347–358, 2008.

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 genet*, 2(12):e190, 2006.

Desislava Petkova, John Novembre, and Matthew Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *bioRxiv*, November 2014. doi: 10.1101/011809. URL `http://biorxiv.org/content/early/2014/11/26/011809`.

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.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Sohini Ramachandran, Omkar Deshpande, Charles C. Roseman, Noah A. Rosenberg, Marcus W. Feldman, and L. Luca Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–15947, 2005. doi: 10.1073/pnas.0507611102. URL `http://www.pnas.org/content/102/44/15942.abstract`.

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`.

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.

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

S Wright. Isolation by distance. *Genetics*, 28(2):114–138, March 1943. URL `http://www.genetics.org/cgi/reprint/28/2/114`.

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`.