Local PCA Shows How the Effect of Population Structure Differs Along the Genome

Han Li[1], Peter Ralph[1,2,3,*]

1 Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA , USA
2 Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA
3 Department of Mathematics, University of Oregon, Eugene, OR, USA

*** plr@uoregon.edu**

Abstract

Population structure leads to systematic patterns in measures of mean relatedness between individuals in large genomic datasets, which are often discovered and visualized using dimension reduction techniques such as principal component analysis (PCA). Mean relatedness is an average of the relationships across locus-specific genealogical trees, which can be strongly affected on intermediate genomic scales by linked selection and other factors, We show how to use local principal components analysis to describe this meso-scale heterogeneity in patterns of relatedness, and apply the method to genomic data from three species, finding in each that the effect of population structure can vary substantially across only a few megabases. In a global human dataset, localized heterogeneity is likely explained by polymorphic chromosomal inversions. In a range-wide dataset of *Medicago truncatula*, factors that produce heterogeneity are shared between chromosomes, correlate with local gene density, and may be caused by background selection or local adaptation. In a dataset of primarily African *Drosophila melanogaster*, large-scale heterogeneity across each chromosome arm is explained by known chromosomal inversions thought to be under recent selection, and after removing samples carrying inversions, remaining heterogeneity is correlated with recombination rate and gene density, again suggesting a role for linked selection. The visualization method provides a flexible new way to discover biological drivers of genetic variation, and its application to data highlights the strong effects that linked selection and chromosomal inversions can have on observed patterns of genetic variation.

1 Introduction

Wright [1] defined *population structure* to encompass "such matters as numbers, composition by age and sex, and state of subdivision", where "subdivision" refers to restricted migration between subpopulations. The phrase is also commonly used to refer to the genetic patterns that result from this process, as for instance reduced mean relatedness between individuals from distinct populations. However, it is not necessarily clear what aspects of demography should be included in the concept. For instance, Blair [2] defines *population structure* to be the sum total of "such factors as size of breeding populations, periodic fluctuation of population size, sex ratio, activity range and *differential survival of progeny*" (emphasis added). The definition is similar to Wright's, but differs in including the effects of natural selection. On closer examination, incorporating differential survival or fecundity makes the concept less clear: should a randomly mating population consisting of two types that are partially reproductively isolated from each other be said to show population structure or not? Whatever the definition, it is clear that due to natural selection, the effects of population structure – the *realized* patterns of genetic relatedness – differ depending on which portion of the genome is being considered. For instance, strongly locally adapted alleles of a gene will be selected against in migrants to different habitats, increasing genetic differentiation between populations near to this gene. Similarly, newly adaptive alleles spread first in local populations. These observations motivate many methods to search for genetic loci under selection, as for example in [3], [4], and [5].

These realized patterns of genetic relatedness summarize the shapes of the genealogical trees at each location along the genome. Since these trees vary along the genome, so does relatedness, but averaging over sufficiently many trees we hope to get a stable estimate that doesn't depend much on the genetic markers chosen. This is not guaranteed: for instance, relatedness 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 [6, 7, 8]. Indeed, many species show chromosome-scale variation in diversity and divergence (e.g., [9]); species phylogenies can differ along the genome due to incomplete lineage sorting, adaptive introgression and/or local adaptation (e.g., [10, 11, 12, 13, 14]); and theoretical expectations predict that geographic patterns of relatedness should depend on selection [15].

Patterns in genome-wide relatedness are often summarized by applying principal components analysis (PCA, [16]) to the genetic covariance matrix, as pioneered by [17]. The results of PCA can be related to the genealogical history of the samples, such as time to most recent common ancestor and migration rate between populations [18, 19], and sometimes produce "maps" of population structure that reflect the samples' geographic origin distorted by rates of gene flow [20].

Modeling such "background" kinship between samples is essential to genome-wide association studies (GWAS, [21, 22]), and so understanding variation in kinship along the genome could lead to more generally powerful methods, and may be essential for doing

GWAS in species with substantial heterogeneity in realized patterns of mean relatedness along the genome.

PCA has been applied to genomic windows in methods to infer tracts of local ancestry in recently admixed populations [23, 24], and to identify putative chromosomal inversions [25].

A note on nomenclature: In this work we describe variation in patterns of relatedness using local PCA, where "local" refers to proximity along the genome. A number of general methods for dimensionality reduction also use a strategy of "local PCA" (e.g., [26, 27, 28, 29]), performing PCA not on the entire dataset but instead on subsets of observations, providing local pictures which are then stitched back together to give a global picture. At first sight, this differs from our method in that we restrict to subsets of *variables* instead of subsets of observations. However, if we flip perspectives and think of each genetic variant as an observation, our method shares common threads, although our method does not subsequently use adjacency along the genome, as we aim to identify similar regions that may be distant.

It is common to describe variation along the genome of simple statistics such as F_{ST} and to interpret the results in terms of the action of selection (e.g., [30, 11]). However, a given pattern (e.g., valleys of F_{ST}) can be caused by more than one biological process [31, 32], which in retrospect is unsuprising given that we are using a single statistic to describe a complex process. It is also common to use methods such as PCA to visualize large-scale patterns in mean genome-wide relatedness. In this paper we show if and how patterns of mean relatedness vary systematically along the genome, in a way particularly suited to large samples from geographically distributed populations. Geographic population structure sets the stage by establishing "background" patterns of relatedness; our method then describes how this structure is affected by selection and other factors. Our aim is not to identify outlier loci, but rather to describe larger-scale variation shared by many parts of the genome; correlation of this variation with known genomic features can then be used to uncover its source.

2 Materials and Methods

As depicted in Fig 1, the general steps to the method are: (1) divide the genome into windows, (2) summarize the patterns of relatedness in each window, (3) measure dissimilarity in relatedness between each pair of windows, (4) visualize the resulting dissimilarity matrix using multidimensional scaling (MDS), and (5) combine similar windows to more accurately visualize local effects of population structure using PCA.

2.1 PCA in genomic windows

To begin, we first recoded sampled genotypes as numeric matrices in the usual manner, by recording the number of nonreference alleles seen at each locus for each sample. We then divided the genome into contiguous segments ("windows") and applied principal component analysis (PCA) as described in [19] separately to the submatrices that corresponded to each window. The choice of window length entails a tradeoff between signal and noise, since shorter windows allow better resolution along the genome but provide less precise estimates of relatedness. A method for choosing a window length to balance these considerations is given in Appendix A. Precisely, denote by Z the $L \times N$ recoded genotype matrix for a given window (L is the number of SNPs and N is the sample size), and by $\overline{Z_s}$ the mean of non-missing entries for allele s, so that $\overline{Z_s} = \frac{1}{n_s} \sum_j Z_{sj}$, where the sum is over the n_s nonmissing genotypes. We first compute the mean-centered matrix X, as $X_{si} = Z_{si} - \overline{Z_s}$, and preserving missingness. (This mean-centering makes the result not depend on the choice of reference allele, exactly if there is no missing data, and approximately otherwise.) Next, we find the covariance matrix of X, denoted C, as $C_{ij} = \frac{1}{m_{ij}-1} \sum_s X_{si} X_{sj} - \frac{1}{m_{ij}(m_{ij}-1)} (\sum_s X_{si})(\sum_s X_{sj})$, where all sums are over the m_{ij} sites where both sample i and sample j have nonmissing genotypes. The principal components are the eigenvectors of C, normalized to have Euclidean length equal to one, and ordered by magnitude of the eigenvalues.

The top 2–5 principal components are generally good summaries of population structure; for ease of visualization we usually only use the first two (referred to as $PC1$ and $PC2$), and check that results hold using more. 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. (Note, however, that our measure of dissimilarity between windows does not depend on PC ordering.)

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

2.2 Similarity of patterns of relatedness between windows

We think of the local effects of population structure as being summarized by *relative* position of the samples in the space defined by the top principal components. However, we do not compare patterns of relatedness of different genomic regions by directly comparing the PCs, since rotations or reflections of these imply identical patterns of relatedness. Instead, we compare the low-dimensional approximations of the local covariance matrices obtained using the top k PCs, which is invariant under ordering of the PCs , reflections, and rotations and yet contains all other information about the PCs. (For results shown here, we use $k = 2$; results using larger numbers of PCs were nearly identical.) Furthermore, to remove the effect of artifacts such as mutation rate variation, we also rescale each approximate covariance matrix to be of similar size (precisely, so that the underlying data matrix has trace norm equal to one).

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

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

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

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

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

2.3 Visualization of results

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

We then locate variation in patterns of relatedness along the genome by choosing collections of windows that are nearby in MDS coordinates, and map their positions along the genome. A visualization of the effects of population structure across the entire collection is formed by extracting the corresponding genomic regions and performing PCA on all, aggregated, regions.

2.4 Datasets

We applied the method to genomic datasets with good geographic sampling: 380 African *Drosophila melanogaster* from the Drosophila Genome Nexus [33], a worldwide dataset of humans, 3,965 humans from several locations worldwide from the POPRES dataset [34], and 263 *Medicago truncatula* from 24 countries around the Mediterranean basin a range-wide dataset of the partially selfing weedy annual plant from the *Medicago truncatula* Hapmap Project [35], as summarized in Table 1.

***Drosophila melanogaster*:** We used whole-genome sequencing data from the Drosophila Genome Nexus (`http://www.johnpool.net/genomes.html`, [33]), consisting of the Drosophila Population Genomics Project phases 1–3 [9, 36], and additional African genomes [33]. After removing 20 genomes with more than 8% missing data, we were left with 380 samples from 16 countries across Africa and Europe. Since the *Drosophila* samples are from inbred lines or haploid embryos, we treat the samples as haploid when recoding; regions with residual heterozygosity were marked as missing in the original dataset; we also removed positions with more than 20% missing data. Each chromosome arm we investigated (X, 2L, 2R, 3L, and 3R) has 2–3 million SNPs; PCA plots for each arm are shown in Fig S2.

Human: We also used genomic data from the entire POPRES dataset [34], which has array-derived genotype information for 447,267 SNPs across the 22 autosomes of 3,965 samples in total: 346 African-Americans, 73 Asians, 3,187 Europeans and 359 Indian Asians. Since these data derive from genotyping arrays, the SNP density is much lower than the other datasets, which are each derived from whole genome sequencing. We excluded the sex chromosomes and the mitochondria. PCA plots for each chromosome, separately, are shown in Fig S3.

***Medicago truncatula*:** Finally, we used whole-genome sequencing data from the *Medicago truncatula* Hapmap Project [35], which has 263 samples from 24 countries, primarily distributed around the Mediterranean basin. Each of the 8 chromosomes has 3–5 million SNPs; PCA plots for these are shown in Fig S4. We did not use the mitochondria or chloroplasts.

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

Table 1: Descriptive statistics for each dataset used.

2.5 Data access

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

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

3 Results

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

3.1 Validation

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

3.2 *Drosophila melanogaster*

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

To help visualize how clustered windows with similar patterns of relatedness are along each chromosome arm, we selected three "extreme" windows in the MDS plot and the 5%

of windows that are closest to it in the MDS coordinates, then highlighted these windows' positions along the genome, and created PCA plots for the windows, combined. Representative plots are shown for three groups of windows on each chromosome arm in Fig 2 (groups are shown in color), and in Supplemental Fig S1 (PCA plots). The latter plots are quite different, showing that genomic windows in different regions of the MDS plot indeed show quite different patterns of relatedness.

The most striking variation in patterns of relatedness turns out to be explained by several large inversions that are polymorphic in these samples, discussed in [37] and [9]. To depict this, Fig 3 shows the PCA plots in Fig S1 recolored by the orientation of the inversion for each sample. Taking chromosome arm 2L as an example, the two regions of similar, extreme patterns of relatedness shown in green in the first row of Fig 2 lie directly around the breakpoints of the inversion In(2L)t, and the PCA plots in the first rows of Fig 3 shows that patterns of relatedness here are mostly determined by inversion orientation. The regions shown in purple on chromosome 2L lie near the centromere, and have patterns of relatedness reflective of two axes of variation, seen in Fig S1 and 3, which correspond roughly to latitude within Africa and to degree of cosmopolitan admixture respectively (see [33] for more about admixture in this sample). The regions shown in orange on chromosome 2L mostly lie inside the inversion, and show patterns of relatedness that are a mixture between the other two, as expected due to recombination within the (long) inversion [38]. Similar results are found in other chromosome arms, albeit complicated by the coexistence of more than one polymorphic inversion; however, each breakpoint visibly affects patterns in the MDS coordinates (see vertical lines in Fig 2).

To see how patterns of relatedness vary in the absence of polymorphic inversions, we performed the same analyses after removing, for each chromosome arm, any samples carrying inversions on that arm. In the result, shown in Supplemental Fig S5, the striking peaks associated with inversion breakpoints are gone, and previously smaller-scale variation now dominates the MDS visualization. For instance, the majority of the variation along 3L in Fig 2 is on the left end of the arm, dominated by two large peaks around the inversion breakpoints; there is also a relatively small dip on the right end of the arm (near the centromere). In contrast, Supplemental Fig S5 shows that after removing polymorphic inversions, remaining structure is dominated by the dip near the centromere. Without inversions, variation in patterns of relatedness shown in the MDS plots follows similar patterns to that previously seen in *D. melanogaster* recombination rate and diversity [9, 39]. Indeed, correlations between the recombination rate in each window and the position on the first MDS coordinate are highly significant (Spearman's $\rho = 0.54$, $p < 2 \times 10^{-16}$; Fig 4 and S6). This is consistent with the hypothesis that variation is due to selection, since the strength of linked selection increases with local gene density, measured in units of recombination distance. The number of genes – measured as the number of transcription start and end sites within each window – was not significantly correlated with MDS coordinate ($p = 0.22$).

3.3 Human

As we did for the *Drosophila* data, we applied our method separately to all 22 human autosomes. On each, variation in patterns of relatedness was dominated by a small number of windows having similar patterns of relatedness to each other that differed dramatically from the rest of the chromosome. These may be primarily inversions: outlying windows coincide with three of the six large polymorphic inversions described in [40], notably a particularly large, polymorphic inversion on 8p23 (Fig 5). Similar plots for all chromosomes are shown in Supplementary Fig S7, S8, and S9. PCA plots of many outlying windows show a characteristic trimodal shape (shown for chromosome 8 in Fig S10), presumably distinguishing samples having each of the three diploid genotypes for each inversion orientation (although we do not have data on orientation status). This trimodal shape has been proposed as a method to identify inversions [25], but distinguishing this hypothesis from others, such as regions of low recombination rate, would require additional data.

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

3.4 *Medicago truncatula*

Unlike the other two species, the method applied separately on all eight chromosomes of *Medicago truncatula* showed similar patterns of gradual change in patterns of relatedness across each chromosome, with no indications of chromosome-specific patterns. This consistency suggests that the factor affecting the population structure for each chromosome is the same, as might be caused by varying strengths of linked selection. To verify that variation in the effects of population structure is shared across chromosomes, we applied the method to all chromosomes together. Results for chromosome 3 are shown in Fig 6 and 6, and other chromosomes are similar: across chromosomes, the high values of the first MDS coordinate coincide with the position of the heterochromatic regions surrounding the centromere, which often have lower gene density and may therefore be less subject to linked selection. To verify that this is a possible explanation, we counted the number of genes found in each window using gene models in Mt4.0 from `jcvi.org` [35], which are shown juxtaposed with the first MDS coordinate of each window in Fig 7, and are significantly correlated, as shown in Supplemental Fig S12. (Values shown are the number of start and end positions of each predicted mRNA transcript, divided by two, assigned to the nearest window.) However, other genomic features, such as distance to centromere show roughly the same patterns, so we cannot rule out alternative hypotheses. In particular, fine-scale recombination rate estimates are not available in a form mappable to Mt4.0 coordinates (although those in [41] appear visually similar).

We also found nearly identical results when choosing shorter windows of 1,000 SNPs;

or choosing windows of equal length in base pairs rather than SNPs. Similarly, the results were not substantially changed when using weighted PCA to downweight the large group of Tunisian samples.

4 Discussion

Our investigations have found substantial variation in the patterns of relatedness formed by population structure across the genomes of three diverse species, revealing distinct biological processes driving this variation in each species. More investigation, particularly on more species and datasets, will help to uncover what aspects of species history can explain these differences. With growing appreciation of the heterogeneous effects of selection across the genome, especially the importance of adaptive introgression and hybrid speciation [13, 42, 43, 44, 45], local adaptation [46, 47], and inversion polymorphisms [48, 49], local PCA may prove to be a useful exploratory tool to discover important genomic features.

We now discuss possible implications of this variation in the effects of population structure, the impact of various parameter choices in implementing the method, and possible additional applications.

Chromosomal inversions A major driver of variation in patterns of relatedness 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 [25]; however, the signal left by inversions likely cannot be distinguished from long haplotypes under balancing selection or simply regions of reduced recombination without additional lines of evidence. Inversions show up in our method because across the inverted region, most gene trees share a common split that dates back to the origin of the inversion. However, in many applications, inversions are a nuisance. For instance, SMARTPCA [16] reduces their effect on PCA plots by regressing out the effect of linked SNPs on each other. Removing samples with the less common orientation of each inversion reduced, but did not eliminate, the signal of inversions seen in the *Drosophila melanogaster* dataset, demonstrating that the genomic effects of transiently polymorphic inversions may outlast the inversions themselves.

The effect of selection It seems that the variation in patterns of relatedness we see in the *Medicago truncatula* and *Drosophila melanogaster* datasets must be explained somehow by linked selection. Furthermore, the selection must be affecting many targets across the genome, since we see similar effects across long distances (even distinct chromosomes). For this reason, the most likely candidate may be selection against linked deleterious mutations, known as "background selection" [7, 50]. Informally, background selection reduces the number of potential contributors to the gene pool in regions of the genome with many possible deleterious mutations [51]; for this reason, if it acts in a spatial context, it is expected to induce samples from nearby locations to cluster together more frequently, Therefore, regions of the genome harboring many targets of local adaptation may show similar patterns, since migrant alleles in these regions will be selected against, and so locally gene trees will more closely reflect spatial proximity.

A related possibility is that variation in patterns of relatedness is due to recent admixture between previously separated populations, the effects of which were not uniform across the genome due to selection. For instance, it has been hypothesized that large-scale variation in amount of introgressed Neanderthal DNA along the genome is due to selection against Neanderthal genes, leading to greater introgression in regions of lower gene density [52, 53]. African *Drosophila melanogaster* are known to have a substantial amount of recently introgressed genome from “cosmopolitan” sources; if selection regularly favors genes from one origin, this could lead to substantial variation in patterns of relatedness correlated with local gene density.

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

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

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

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

Applications 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 [22, 54]. Since the effects of population structure is not constant along the genome, this could in principle lead to an inflation of false positives in parts of the genome with stronger population structure than the genome-wide average. A method such as ours might be used to provide a more sensitive correction. 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 would entail a substantial loss of power, and could potentially be misleading.

Acknowledgements

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

Disclosure declaration

The authors declare no conflicts of interest.

References

[1] Wright S. The genetical structure of populations. Annals of Eugenics. 1949;15(1):323–354. doi:10.1111/j.1469-1809.1949.tb02451.x.

[2] Blair AP. Population Structure in Toads. The American Naturalist. 1943;77(773):563–568.

[3] Huerta-Sánchez E, DeGiorgio M, Pagani L, Tarekegn A, Ekong R, Antao T, et al. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Molecular Biology and Evolution. 2013;30(8):1877–1888. doi:10.1093/molbev/mst089.

[4] Martin SH, Moest M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. Genetics. 2016;203(1):525–541. doi:10.1101/042796.

[5] Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Molecular Biology and Evolution. 2015;doi:10.1093/molbev/msv334.

[6] Kim Y, Stephan W. Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. Genetics. 2002;160(2):765–777.

[7] Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–1303.

[8] Barton NH. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci. 2000;355(1403):1553–1562. doi:10.1098/rstb.2000.0716.

[9] Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, et al. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics. 2012;192(2):533–598. doi:10.1534/genetics.112.142018.

[10] Pease JB, Hahn MW. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. Evolution. 2013;67(8):2376–2384. doi:10.1111/evo.12118.

[11] Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, et al. The genomic landscape of species divergence in Ficedula flycatchers. Nature. 2012;491(7426):756–760. doi:10.1038/nature11584.

[12] Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos Trans R Soc Lond B Biol Sci. 2012;367(1587):343–353. doi:10.1098/rstb.2011.0198.

[13] Pool JE. The Mosaic Ancestry of the Drosophila Genetic Reference Panel and the *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness Interactions. Molecular Biology and Evolution. 2015;32(12):3236–3251. doi:10.1101/014837.

[14] Vernot B, Akey JM. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science. 2014;doi:10.1126/science.1245938.

[15] Charlesworth B, Charlesworth D, Barton NH. The effects of genetic and geographic structure on neutral variation. Annual Review of Ecology, Evolution, and Systematics. 2003;34(1):99–125. doi:10.1146/annurev.ecolsys.34.011802.132359.

[16] Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genetics. 2006;2(12):e190. doi:10.1371/journal.pgen.0020190.

[17] Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. Science. 1978;201(4358):786–792.

[18] Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nature genetics. 2008;40(5):646–649.

[19] McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009;5(10):e1000686.

[20] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98–101.

[21] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904–909.

[22] Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statistical Science. 2009;24(4):451–471. doi:10.1214/09-STS307.

[23] Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A. 2010;107(2):786–791. doi:10.1073/pnas.0909559107.

[24] Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol. 2012;84(4):343–364.

[25] Ma J, Amos CI. Investigation of inversion polymorphisms in the human genome using principal components analysis. PLoS One. 2012;7(7). doi:10.1371/journal.pone.0040224.

[26] Manjón JV, Coupé P, Concha L, Buades A, Collins DL, Robles M. Diffusion weighted image denoising using overcomplete local PCA. PloS one. 2013;8(9):e73021.

[27] Kambhatla N, Leen TK. Dimension Reduction by Local Principal Component Analysis. Neural Computation. 1997;9(7):1493–1516. doi:10.1162/neco.1997.9.7.1493.

[28] Weingessel A, Hornik K. Local PCA algorithms. Neural Networks, IEEE Transactions on. 2000;11(6):1242–1250.

[29] Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 2000;290(5500):2323–2326. doi:10.1126/science.290.5500.2323.

[30] Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 2005;3(9). doi:10.1371/journal.pbio.0030285.

[31] Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23(13):3133–3157. doi:10.1111/mec.12796.

[32] Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome Res. 2015;25(11):1656–1665. doi:10.1101/gr.196485.115.

[33] Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The Drosophila genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics. 2015;199(4):1229–1241.

[34] Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet. 2008;83(3):347–358. doi:10.1016/j.ajhg.2008.08.005.

[35] Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, et al. An improved genome release (version Mt4. 0) for the model legume *Medicago truncatula*. BMC genomics. 2014;15(1):1.

[36] Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al. Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. PLoS Genet. 2012;8(12):1–24. doi:10.1371/journal.pgen.1003080.

[37] Corbett-Detig RB, Hartl DL. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. PLoS Genet. 2012;8(12):e1003056.

[38] Guerrero RF, Rousset F, Kirkpatrick M. Coalescent patterns for chromosomal inversions in divergent populations. Philosophical Transactions of the Royal Society B: Biological Sciences. 2011;367(1587):430–438. doi:10.1098/rstb.2011.0246.

[39] Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The *Drosophila melanogaster* Genetic Reference Panel. Nature. 2012;482(7384):173–178.

[40] Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, et al. Characterization of six human disease-associated inversion polymorphisms. Human molecular genetics. 2009;18(14):2555–2566.

[41] Paape T, Zhou P, Branca A, Briskine R, Young N, Tiffin P. Fine-Scale Population Recombination Rates, Hotspots, and Correlates of Recombination in the *Medicago truncatula* Genome. Genome Biology and Evolution. 2012;4(5):726–737. doi:10.1093/gbe/evs046.

[42] Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet. 2014;10(6):e1004410. doi:10.1371/journal.pgen.1004410.

[43] Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The Genomic Signature of Crop-Wild Introgression in Maize. PLoS Genet. 2013;9(5):e1003477. doi:10.1371/journal.pgen.1003477.

[44] Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. Rapid spread of invasive genes into a threatened native species. Proc Natl Acad Sci U S A. 2010;107(8):3606–3610. doi:10.1073/pnas.0911802107.

[45] Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (*Mus musculus*). PLoS Genet. 2012;8(8):e1002891. doi:10.1371/journal.pgen.1002891.

[46] Lenormand T. Gene flow and the limits to natural selection. Trends in Ecology & Evolution. 2002;17(4):183 – 189. doi:DOI: 10.1016/S0169-5347(02)02497-7.

[47] Wang IJ, Bradburd GS. Isolation by environment. Molecular Ecology. 2014;23(23):5649–5662. doi:10.1111/mec.12938.

[48] Kirkpatrick M, Barrett B. Chromosome inversions, adaptive cassettes and the evolution of species' ranges. Molecular Ecology. 2015;doi:10.1111/mec.13074.

[49] Kirkpatrick M. How and why chromosome inversions evolve. PLoS Biol. 2010;8(9). doi:10.1371/journal.pbio.1000501.

[50] Charlesworth B. Background Selection 20 Years on: The Wilhelmine E. Key 2012 Invitational Lecture. Journal of Heredity. 2013;104(2):161–171. doi:10.1093/jhered/ess136.

[51] Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605–1617.

[52] Harris K, Nielsen R. The Genetic Cost of Neanderthal Introgression. Genetics. 2016;203(2):881–891.

[53] Juric I, Aeschbacher S, Coop G. The Strength of Selection Against Neanderthal Introgression. bioRxiv. 2016;doi:10.1101/030148.

[54] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100–106.

[55] Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. *Drosophila melanogaster* recombination rate calculator. Gene. 2010;463(1–2):18 – 20. doi:http://dx.doi.org/10.1016/j.gene.2010.04.015.

[56] Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 2012;8(10). doi:10.1371/journal.pgen.1002905.

[57] Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics; 1982. Available from: `http://epubs.siam.org/doi/abs/10.1137/1.9781611970319`.

[58] Busing FMTA, Meijer E, Van Der Leeden R. Delete-m Jackknife for Unequal m. Statistics and Computing. 1999;9(1):3–8. doi:10.1023/A:1008800423698.

[59] Engelhardt BE, Stephens M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 2010;6(9). doi:10.1371/journal.pgen.1001117.

[60] Qiu Y, Mei J. RSpectra: Solvers for Large Scale Eigenvalue and SVD Problems; 2016. Available from: `https://CRAN.R-project.org/package=RSpectra`.

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

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

Fig. 3: PCA plots for the three sets of genomic windows colored in Fig 2, on each chromosome arm of *Drosophila melanogaster*. In all plots, each point represents a sample. The first column shows the combined PCA plot for windows whose points are colored green in Fig 2; the second is for orange windows; and third is for purple windows. In each, samples are colored by orientation of the polymorphic inversions In(2L)t, In(2R)NS, In(3L)OK, In(3R)K and In(1)A respectively (data from [33]). In each "INV" denotes an inverted genotype, "ST" denotes the standard orientation, and "N" denotes unknown.

Fig. 4: The effects of population structure without inversions is correlated to recombination rate in *Drosophila melanogaster*. The first plot (in red) shows the first MDS coordinate along the genome for windows of 10,000 SNPs, obtained after removing samples with inversions. (A plot analogous to Fig 2 is shown in Supplemental Fig S5.) The second plot (in blue) shows local average recombination rates in cM/Mbp, obtained as midpoint estimates for 100Kbp windows from the Drosophila recombination rate calculator [55] release 5, using rates from [56]. The third plot (in black) shows the number of genes' transcription start and end sites within each 100Kbp window, divided by two. Transcription start and end sites were obtained from the RefGene table from the UCSC browser. The histone gene cluster on chromosome arm 2L is excluded.

Fig. 5: Variation in structure between windows on human chromosomes 8, 15, and 17. Each point in each plot represents a window. The first column shows the MDS visualization of relationships between windows; the second and third columns show the two MDS coordinates of each window against its position (midpoint) along the chromosome. Rows, from top to bottom show chromosomes 8, 15, and 17. The vertical red lines show the breakpoints of known inversions from [40].

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

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

A Choosing window length

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

Since we summarize patterns of relatedness 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. The definition of eigenvectors does not specify their sign, and so when comparing between windows we choose signs to best match each other: after choosing $PC1_1$, for instance, if u is the first eigenvector obtained from the covariance matrix for window j, then we next choose $PC1_j = \pm u$, where the sign is chosen according to which of $\|PC1_1 - u\|$ or $\|PC1_1 + u\|$ is smaller.

After doing this, the mean variance across windows is

$$\sigma^2_{\text{signal}} = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{L} \sum_{i=1}^{L} \left(PC1_{ij} - \overline{PC1_j} \right)^2 ,$$

where $PC1_{ij}$ is the position of the i^{th} individual on $PC1$ in window j, and $\overline{PC1_j} = (1/N) \sum_{j=1}^{N} PC1_{ij}$. We estimate the standard error for each $PC1_{ij}$ using the block jackknife [57, 58]: we divide the j^{th} window into 10 equal-sized pieces, and let $PC1_{ij,k}$ denote the first principal component of this region found after removing the k^{th} piece; then the estimate of the squared standard error is $\sigma^2_{ij} = \frac{9}{10} \sum_{k=1}^{10} (PC1_{ij,k} - \frac{1}{10} \sum_{\ell=1}^{10} PC1_{ij,\ell})^2$. Averaging over samples and windows,

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

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

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

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

B Weighted PCA

Principal components analysis can be thought of as finding a good low-dimensional matrix factorization [59] that well-approximates the original data in the least-squares sense: if C is the $N \times N$ genetic covariance matrix, then to find the top k principal components, we find an orthogonal $N \times k$ matrix U, and a $k \times k$ diagonal matrix Λ with diagonal entries $\Lambda_{ii} = \lambda_i$ to minimize

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

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

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

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

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

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

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

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

In our R implementation we use the Spectra library [60] to find only the top k eigenvectors.

List of Figures

C Supplementary Tables

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

[[2]]

MDS1	10000 SNPs, 2 PCs	1000 SNPs, 2 PCs	10000 SNPs, 5 PCs	100000bp, 2 PCs	10
10000 SNPs, 2 PCs	1.00	0.54	0.93	0.87	
1000 SNPs, 2 PCs	0.82	1.00	0.76	0.83	
10000 SNPs, 5 PCs	0.93	0.50	1.00	0.83	
100000bp, 2 PCs	0.87	0.59	0.84	1.00	
10000bp, 2 PCs	0.83	0.92	0.77	0.84	

MDS2	10000snp, 2 PCs	1000snp, 2 PCs	10000snp, 5 PCs	100000bp, 2 PCs	10000bp, 2 P
10000snp, 2 PCs	1.00	0.54	0.93	0.87	0.
1000snp, 2 PCs	0.82	1.00	0.76	0.83	0.
10000snp, 5 PCs	0.93	0.50	1.00	0.83	0.
100000bp, 2 PCs	0.87	0.59	0.84	1.00	0.
10000bp, 2 PCs	0.83	0.92	0.77	0.84	1.

Table S2: C

orrelations between MDS coordinates of genomic regions between runs with different parameter values. To produce these, we first ran the algorithm with the specified window size and number of PCs (k in equation *XXX*) on the full *Medicago truncatula* dataset. Then to obtain the correlation between results obtained from parameters A in the row of the matrix above and parameters B in the column of the matrix above, we mapped the windows of B to those of A by averaging MDS coordinates of any windows of B whose midpoints lay in the corresponding window of A; we then computed the correlation between the MDS coordinates of A and the averaged MDS coordinates of B. This is not a symmetric operation, so these matrices are not symmetric. As expected, parameter values with smaller windows produce noisier estimates. Plots of MDS values along the genome are visually nearly identical for parameter sets having similar window sizes – full reports are available as supplementary material on Data Dryad *XXX*.

D Supplementary Figures

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

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

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

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

Fig. S5: Variation in structure for windows of 1,000 SNPs across *Drosophila melanogaster* chromosome arms: without inversions. As in Fig 2, but after omitting for each chromosome arm individuals carrying the less frequent orientation of any inversions on that chromosome arm. The values differ from those in 4 in the window size used and that some MDS values were inverted (but relative orientation is meaningless as chromosome arms were run separately, unlike for *Medicago*). In all plots, each point represents one window along the genome. The first column shows the MDS visualization of relationships between windows, and the second and third columns show the midpoint of each window against the two MDS coordinates; rows correspond to chromosome arms. Colors are consistent for plots in each row. Vertical lines show the breakpoints of known polymorphic inversions.

Fig. S6: Recombination rate, and the effects of population structure for *Drosophila melanogaster*: this shows the first MDS coordinate and recombination rate (in cM/Mbp), as in Fig 4, against each other. Since the windows underlying estimates of Fig 4 do not coincide, to obtain correlations we divided the genome into 100Kbp bins, and for each variable (recombination rate and MDS coordinate 1) averaged the values of each overlapping bin with weight proportional to the proportion of overlap. The correlation coefficient and p-values for each linear regression are as follows: 2L: correlation = 0.52, $r^2 = 0.27$; 2R: correlation = 0.43, $r^2 = 0.18$; 3L: correlation = 0.47, $r^2 = 0.21$; 3R: correlation = 0.46, $r^2 = 0.21$; X: correlation = 0.50, $r^2 = 0.24$.

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

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

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

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

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

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

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

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

Resubmission Cover Letter
Genetics

Han Li
and Peter Ralph
December 21, 2017

To the Editor(s) –

We are pleased to submit a revision of our manuscript,

Sincerely,

Han Li and Peter Ralph

Reviewer AE:

> Please understand that incremental changes will not be sufficient. Adding simulations to strengthen key claims will be necessary, particularly addressing the impacts of mutation rate and recombination rate variation with more depth, the concern regarding PC switching (Reviewer 1), and the concern regarding the impacts of variation in missingness by sub-population (Reviewer 2).

Thanks for the positive feedback and the useful suggestions. We agree that more extensive exploration using simulations would help bolster understanding of the method, and have now done so. This took a substantial amount of work, because genome-scale forwards-time simulations with a large number of loci under selection is at or beyond the current limits of computation, depending on the number of individuals simulated.

Reviewer 1:

> The paper is generally well written and clear; it addresses an important problem, and clearly makes some progress on it. However, it suffers from having no grounding in either theory or empirical demonstration that it really can find the structures that are claimed. I find the arguments that it finds inversions compelling, though not watertight, and I am not yet convinced that it is finding ubiquitous background selection. To make this claim, significant extra work is required.
>
> In short, the approach is interesting but not sufficiently explored to produce compelling evidence for the implications that are claimed. Putting a large amount of effort into simulations may alleviate these concerns somewhat.
>
> Specific points: What does this method find? I'm concerned about: (a) variation in the recombination rate and (b) variation in the mutation rate, creating spurious structure.
>
> The first possibility is that massively varying information quantity within windows could lead to a small number of such windows having their orientation reversed: that is, PC1 becomes PC2 and vice versa. (Or PC2 and PC3 could switch). This would lead to such windows having unusual properties and hence appearing as evidence of an inversion.
>
> I do agree with the authors that significant outliers would be found at inversions. However, even if the PC switching does not occur, or the model could handle it, the evidence for selection is weaker. If the two types of variation described above exist, with no selection, I would still expect a "continuous triangle" of results (as seen left of Fig 2, top left of Fig 6) with extrema described by windows

with the most information, and points placed at different extremum having low recombination rate (because by chance, these will get an approximately fixed local tree, corresponding on average to the genome-wide population structure).

Addressing this is likely quite hard, though the authors may be able to think of something that separates these effects from selection.

(1.1) *. . . variation in the recombination rate . . . creating spurious structure.*

Reply: We address this in two ways. This point is addressed by comparing results with windows of different types – windows of equal length in bp (or in SNPs) have different lengths in cM; since these different choices show nearly identical patterns, recombination rate variation cannot be driving the results.

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

Reply:

(1.3) PC switching *. . . could lead to a small number of such windows having their orientation reversed: that is, PC1 becomes PC2 and vice versa. (Or PC2 and PC3 could switch).*

Reply: This is a natural concern. However, the only point at which we compare PCs in a way that could be sensitive to ordering is in determining the window size – in computing the distance between windows we use a measure which is invariant under ordering. We have made this more clear by moving the note about flipping signs of PCs to the appendix on window choice (p. 18, l. 590) and added more explicit notes about this to (p. 4, l. 129) and (p. 4, l. 143).

(1.4) p6 *"here, we use k=2..." - you have to show that $k > 2$ is the same.*

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

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

(1.7) *Using 2 PCs is common practice: only if this is the end of an analysis and the PCA was done for visualisation. Here you are using it for something so should keep all the relavant PCs.*

Reply: This is a good point; the question is which the "relevant" PCs are. [18] showed that under isolation by distance, the top two PCs should reflect the two-dimensional nature of the range, and higher PCs are generally much less interpretable; we used $k = 2$ with this in mind. We have changed this sentence (p. 4, l. 126).

(1.8) *I'm surprised that PCAdmix isn't referenced. It is using a very similar method, albeit with different goals. In particular, the approach of placing all points into a single, genome-wide PC space solves many of the problems that this approach has (though I agree there may be benefits to the approach described here)*

Reply: Good point: we now reference this work (p. 3, l. 75).

Reviewer 2:

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

(2.1) Related work: *The authors provide an introduction of the main concepts, as well as some intuition of what the method is doing and how, but I found comparison to previous approaches to be somewhat missing. To some extent, this is due to the fact that the main goal of their analysis is somewhat vaguely "finding heterogeneity", which leads to the applications of detecting chromosomal inversions and evidence for background selection. It would help to have a well defined set of hypotheses, test the method's accuracy using simulation (see next comment), and compare to previous efforts in similar domains.*

Reply: First: we think that "finding heterogeneity" is in fact a well-defined goal, although it was not that well-defined in the paper; we have hopefully improved on this in the Introduction (p. 3, l. 98). Expanding a bit more: We strongly agree that methods that seek to test well-defined hypotheses are extremely useful and powerful. We also feel that methods for visualization and exploration are also useful – a primary example here being PCA. If PCA is useful – and we think that it is – then it should be important to also know how much the thing that PCA is summarizing varies along the genome, in the same way that knowing the mean of some quantity in a population is only of limited usefulness without also knowing the corresponding populaion variance.

(2.2) Validation: *In several occasions, the authors seem to introduce a potential problem in their approach, and provide a solution to it. This is generally rather intuitive, but it would really help to have simulations of some sort to show that the issue arises and leads to a problem, and that their approach does address the specific problem.*

(2.3) *The use of weighted PCA to cope with unbalanced sample size could be better demonstrated. Although the current explanation makes intuitive sense, this approach does* *not seem to be used in previous work. The authors could design a simulation that supports their approach.*

(2.4) *It is conceivable that some subpopulations will have more missingness in some windows. That may skew the resulting PCs by selecting different sample sizes for the different windows (as discussed in Appendix B) . This could distort the PCs, so that variation reflects underlying variation in missingness. Would be good to discuss this potential issue and provide simulations.*

(2.5) Appendix A: *when using jackknife to estimate variance, each window is being divided in 10 "independent" resampling units. Due to LD, these 10 blocks are likely correlated, which would bias the estimates of variance. This is probably not a problem because both signal and noise could be equally biased, but the authors may want to consider this potential issue. I wonder if the correlation with recombination rate may be partially explained by this.*

(2.6) *Is it possible to explain the results of Figure 6 just considering neutral variation in local ancestry due to recent admixture? This may explain why ancestry seems to explain a fair amount of variance in the lower plots of Fig 6. Local PCA has been previously used by others to detect local ancestry blocks, e.g. see the PCAdmix approach by Brisbin et al. The authors discuss the possibility that admixture is driving the differentiation, but do not test whether their observations agree with neutrality.*

Reply: This is a good point, but XXX We now cite PCAdmix (p. 3, l. 75).

(2.7) *"to remove the effect of artifacts such as mutation rate variation, we also rescale each approximate covariance matrix to be of similar size (precisely, so that the underlying data matrix has trace norm equal to one". This potential issue is a bit unclear to me, since I would expect that scaling the volume of local trees would not result in changed distances in PC space. Perhaps the authors could show via simulations that this creates a problem, and that the normalization addresses it.*

(2.8) Figure 7: *are MDS coordinates correlated with recombination rates in this case?*

Reply: We made a stab at checking this, and obtained the best version to date of the Medicago recombination map from Tim Paape and Peter Tiffin. There are two versions: a very coarse physical map, and a fine-scale map estimated using LDhat. However, both are on version 3.0 of the assembly, while all other coordinates (sequencing data; gene annotations) are in version 4.0. Furthermore, as Peter Tiffin told us, "apparently there are no files that translate Mt3.0 to Mt4.0 locations (yes, seems a bit silly)." There is a liftOver chain file for translating 3.5 to 4.0, and "the differences in the Mt3.0 and Mt3.5 assemblies are, however, apparently relatively minor". On this basis, we produced the desired figure assuming that Mt3.0 coordinates are the same as Mt3.5 coordinates, included to satisfy the reviewers' curiosity:

However, given uncertainties in this mapping, the relatively poor match of window sizes, the large number of unmappable windows, and the nature of the recombination data (produced with LDhat, not with actual observations of recombiations), we decided not to include this (but have provided a note, (p. 9, l. 313)).

(2.9) Application: *is what the authors seem to be proposing not already accounted for by linear mixed model association approaches? If not, this should be clarified. Either way, this paragraph could be dropped.*

(2.10) Introduction: *"it is not necessarily clear what aspects of demography should be included in the concept." I find it a bit weird to describe selection as an "aspect of demography". Although it could be seen as such within a coalescent framework, that seems to be just a useful representation. The authors may consider rewording'.*

(2.11) *Paragraph starting in "Since the definition...". The notation is a bit unclear. Please check that it is clear which PC the text refers to.*

(2.12) *Would the authors be able to provide a sense for the directionality of effects in Figure 4? It would be interesting if the authors tried to further characterize regions that are similar due to higher recombination rates. E.g. is there more/less density of polymorphisms in these regions?*

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

(2.14) *Typo in abstract, line 6 ", We show" -¿ ". We show".*

(2.15) *Typo: end of introduction "an visualization". The whole sentence is a bit weird. The authors just stated focus is on clustering, not on looking for outliers, but what does it mean that "we allow ourselves to be surprised by unexpected signals in the data"?*

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

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

(2.18) *Table 1 legend seems a bit redundant. Columns are self-explanatory.*

Reply: Good point; we've cut this down.

(2.19) *It would help to have numbered lines and references.*

Reply: We greatly prefer named references rather than numbers, but the LaTeXsource code is available at `https://github.com/petrelharp/local_pca` (run 'make local_pca_paper.pdf' in the 'writeup/' directory) – the reviewer is welcome to change the formatting and recompile.