Using Local PCA to Investigate Population Structure Along the Genome

Han Li, Peter Ralph

March 30, 2016

1 Introduction

Should cite something, probably Astle & Balding in this paragraph. (Try to make these next three sentences into one sentence.) 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. *(This definition is not quite right: two alleles are* always *inherited from their most recent common ancestor.)* It could be estimated from given pedigree or from genome-wide covariances of genotype markers. *(Estimating the "kinship matrix" from pedigrees, or from covariances, are two different things.)* It is well-known that for kinship matrix, actual relatednesses have a lot of noise about the expected value, *(explain what the expected value is)* and depend on where on the genome you look; this is why scans for selective sweeps work. *(do you think this last comment about scans for selective sweeps needs explanation?)* Populations are often structured in some way while there are systematic genetic variation between populations. *(Instead, how about: Structure in the kinship matrix, such as a correlation between geographic proximity and relatedness, is often referred to as "population structure".)*

About 37 years ago, Menozzi et al. (1978) first applied principal component analysis

(PCA) in population genetics to construct maps summarizing genetic variation *(instead, "summarizing population structure"?)* (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. *(We can call this the "genetic covariance matrix".)* The results of PCA can be directly related to the underlying genealogical history of the samples, such as *mean* 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 samples 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 *the* 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. *(I don't think it makes sense to say that alleles are close in a gene tree.)* 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. *(Add citations for the following.)* 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. *(I don't think noise affects population structure, since I'd define population structure to be something like the mean, and that noise just obscures population structure.) (Should mention (and cite) why the effects of introgression might be different along the genome.)* 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.

(Replace "this" with what you mean by "this" in the next sentence.) 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 *and* measured the pairwise Euclidean distance between those matrices. We use multidimensional scaling *(MDS)* 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.

(These two sentences need more detail. How does what we're doing differ from them? Especially, say, chromopainter, argweaver, and http://www.ncbi.nlm.nih.gov/pubmed/26945783? What does local PCA mean in other places, or at lesat, what other fields is it used in?) 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). The term "local PCA" is also used in some other region for different meanings. (Kambhatla and Leen 1997; Manjón et al. 2013; Weingessel and Hornik 2000)

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).

Separate paragraphs for each dataset. Say how many sites, what kind of data, etcetera.

For human, we use SNP chip data from POPRES Nelson et al. (2008). There are 3965 samples in total, (346 African-Americas; 73 Asians; 3187 Europeans; 359 Indian Asians), and the 22 autosomes together have 447267 SNPs in this dataset. We use the allele that has highest frequency in the samples as the reference allele for each position. If an allele is same with the reference allele, we recode it as 0; if an allele is different from the reference allele, we recode it as 1; for positions that have missing data, we recode them as NA. Since human genome is diploid, we add the value for the two alleles and then one chromosome will eventually be recode to a sequence consisting of 0,1,2 (and NA). Then the genome data for human is recoded to a matrix that has 3965 columns, each column for an individual's genotype. We process the data separately for the 22 autosomes in human.

For *Drosophila*, we use the sequencing data from DPGP and John pool's lab, which together has 380 samples from 16 countries. We first process the sequencing data (from DPGP and John pool's lab) to SNP data by eliminating the positions that have all the same alleles. Each chromosome arms we investigated (Chr2L, Chr2R, Chr3L, Chr3R, ChrX) 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% NAs and positions with more than 20%. The cutoff points 8% and 20% are determined from the corresponding distributions of NAs in samples and at positions. After we got the SNP data for *Drosophila*, we recode it *Drosophila*to matrices with 0,1,2 (and NA) for each chromosome arms (Chr2L, Chr2R, Chr3L, Chr3R, ChrX) similar to the process for human genome. Since the *Drosophila* samples here are all homogenous, the matrices are indeed consisting of 0,1 (and NA).

For *Medicago truncatula*, we use the SNP data from Medicago Hapmap. It has 263 samples from 24 countries. We recode it to matrices with 0,1,2 (and NA) for each of the 8 chromosomes. Each chromosome has 3-5 million SNPs

3.2 Cut each genome into windows

this describes PCA, not cutting into windows? and, the next section? We cut each recoded matrix into sub matrixes by that have the same columns but fewer rows than the original matrix. Then apply Principal Component Analysis (PCA) on each window. Here's a brief summary about how is PCA carried out on genomic data (McVean 2009). Starting with the recoded genotype matrix Z, where Z is a L*N matrix (L is SNP number ; N is sample size), then zero-center the matrix Z to X to make the data rows have equal variance. Get the covariance matrix of X (denote as matrix C) and compute the eigenvectors and eigenvalues of the covariance matrix C. The ith principal component is the ith eigenvector of C. *(put equations into the text in the right place)*

$$X_{si} = Z_{si} - \frac{1}{n}\sum_{j=1}^{n} Z_{sj} \tag{1}$$

$$C = \frac{1}{n-1} X X^T \tag{2}$$

3.2.1 Genomic PCA on windows

PCA on each window is getting the eigenvectors of covariance matrix of the cut recoded matrix. PCA plots (generally using PC2 against PC1) can show population structure, which describes how different samples are related. We want to investigate the variation of population structure along the genome. So we apply PCA on each window to check the population structure along the genome.

3.2.2 Choose window length

The window length should neither be too long or too short. The longer the windows, the more accurate is the estimate of population structure in that window. However, for better resolution, we need to find a length that is reasonably short. 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 as mean variance of the first principal component for each window. The variance inside a window is estimated using the jackknife block when cutting the window into 10 equal size smaller windows Efron and Efron (1982). Table 1 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* separately.

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

Table 1: Comparison of variance within a window and that between windows for chromosome arms in *Drosophila*.

3.2.3 Similarity of population structure between windows

We use the first two principal components (PCs) and the corresponding eigenvalues from PCA, and construct a new matrix with the two PCs for each window to stand for the population structure information. 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; λ_{1j} and λ_{2j} are the eigenvalues for the first two PCs for jth window.) M_i and M_j are 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).*

$$\begin{aligned} M_i &= \frac{\lambda_{1i} PC1_i PC1_i^T + \lambda_{2i} PC2_i PC2_i^T}{\lambda_{1i} + \lambda_{2i}} && (3) \\ M_j &= \frac{\lambda_{1j} PC1_j PC1_j^T + \lambda_{2j} PC2_j PC2_j^T}{\lambda_{1j} + \lambda_{2j}} && (4) \end{aligned}$$

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

Define D.

$$V_1 = \sqrt{\frac{\lambda_{1i}}{\lambda_{1i} + \lambda_{2i}}} PC1_i \qquad V_2 = \sqrt{\frac{\lambda_{2i}}{\lambda_{1i} + \lambda_{2i}}} PC2_i \tag{5}$$

$$\begin{aligned} D_{ij} = \Big\{ &(V_1 \cdot V_1)^2 + (V_2 \cdot V_2)^2 + (U_1 \cdot U_1)^2 + (U_2 \cdot U_2)^2 \\ &- 2 \left[(V_1 \cdot U_1)^2 + (V_1 \cdot U_2)^2 + (V_2 \cdot U_1)^2 + (V_2 \cdot U_2)^2 \right] \Big\}^{1/2} \end{aligned} \tag{6}$$

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

3.2.4 Visualize the pairwise distance matrix

To do this, I use Multidimensional scaling (MDS), which is a visualization method commonly applied to distance matrixes. It can reduce the dimensionality of a distance matrix while preserve the distance information between objects as well as possible. The aimed dimension “ M” can be set based on your need. We used M=1 and M=2 in our study. This allowed us to visualize the information in the distance matrix in one or two dimensional relation between windows' population structure.

4 Results

In all these 3 species, PCA plots vary along the genome.

Since PCA plots can show population structure, this shows that the population structure varies along the genome. Each PC plot comes from the covariance matrix of a different section of the genome, which may result from different tree for each position, chromosome inversions, linked selection and so on. We investigate the pattern in each species here separately, combining MDS method and genome features. Here are the MDS results for each species.

4.1 *Drosophila*

We checked the results for chromosome arms Chr2L, Chr2R, Chr3L, Chr3R and ChrX separately. When setting the dimension parameter M=2 in MDS method, that is, plotting the first 2 coordinates reduced from the distance matrix, each plot looks like triangle. (Figure 1a) Since the relative position for each window in the plot shows the relative similarity between windows, it tells us there are 3 extreme types of population structure shown in the 3 peaks of the “triangle”, and other windows are between them, which means

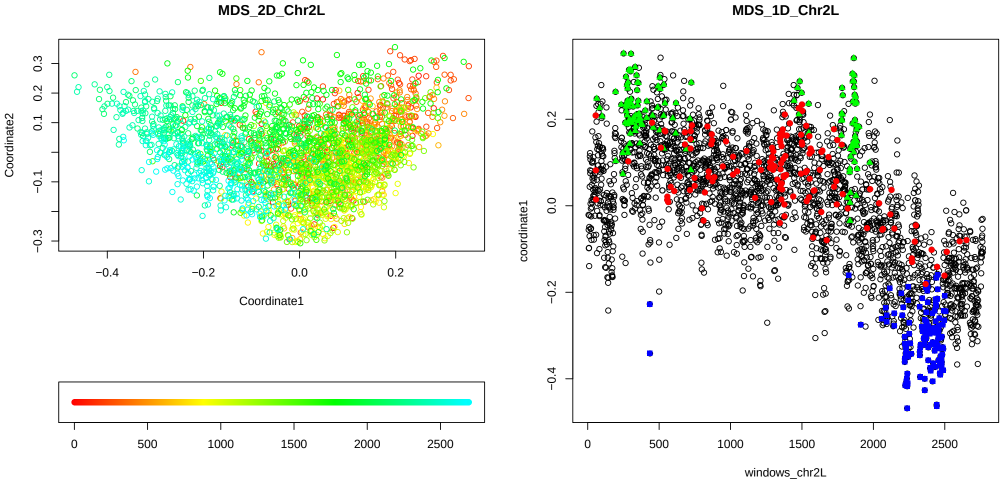

Figure 1: a) Two dimensional MDS plot for chromosome arm Chr2L in *Drosophila*. Each point in the plot stands for a window. The distance between windows shows the similarity of population structure for windows. b) The color corresponding to that in a) and thus shows a relative position for each window in a) along the genome. c) MDS 1D plot for Chr2L in *Drosophila*, using the first coordinate in a) against the window id along the genome. The green, red, blue points are corresponding to the windows in the green, red, blue circled extremes in a). *(I think it's a bit confusing having both MDS1D and MDS2D: can we instead of MDS1D on the right, just plot the first coordinate of MDS2D? And, somehow make the colors agree between the two figures? (not sure how to do this)*

other window's population structure might be a mixture of those extremes. We then want to investigate more information at the extremes.

We pick a window for each extreme, and take out 5% windows that have the smallest distance to it in the original pairwise distance matrix, then combine those windows for each extreme and apply PCA on them. (Fig.2) We can see the obvious difference between their PCA plots. This difference stands for variation of population structure along the genome in *Drosophila* Chr2L.

Use figure labels, like "Figure \ref{fig:mds_chr2L}" instead of "Figure 1".

There's a known inversion in Chr2L in *Drosophila*, In(2L)t, with breakpoints at 2225744bp

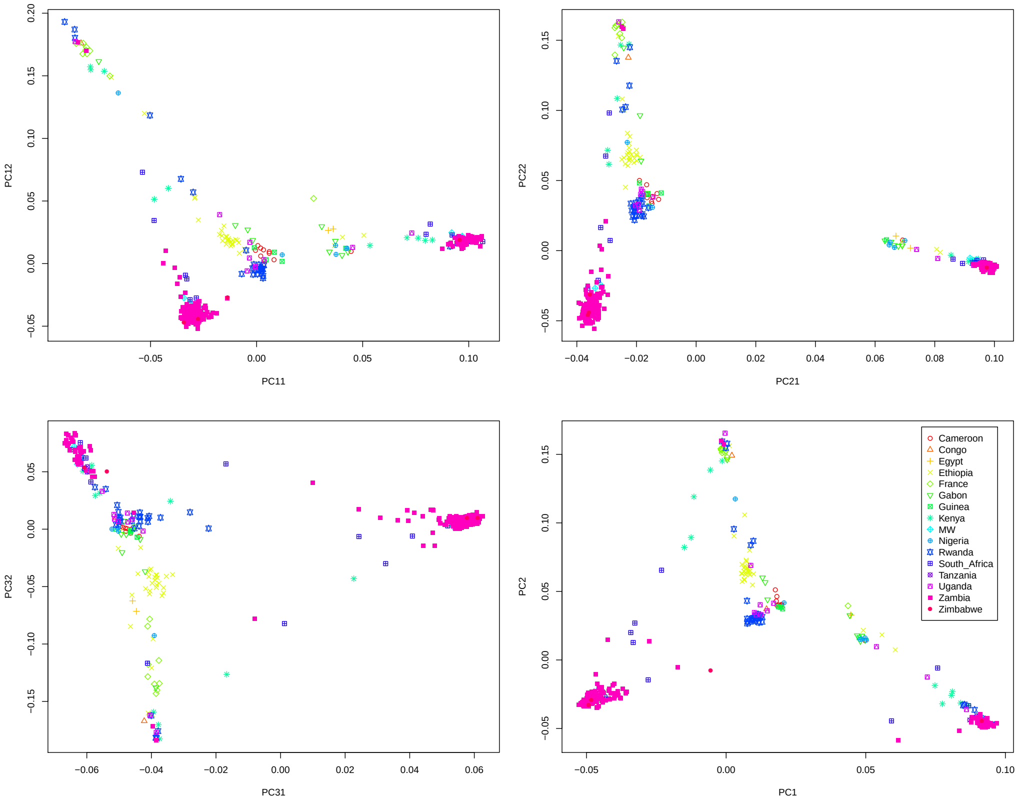

Figure 2: PCA plots for each peak in Figure 1a and for whole Chr2L in *Drosophila*. The color of the square corresponds to the circles of the 3 extremes in Fig.1a. (Also labeled same as 1,2,3)

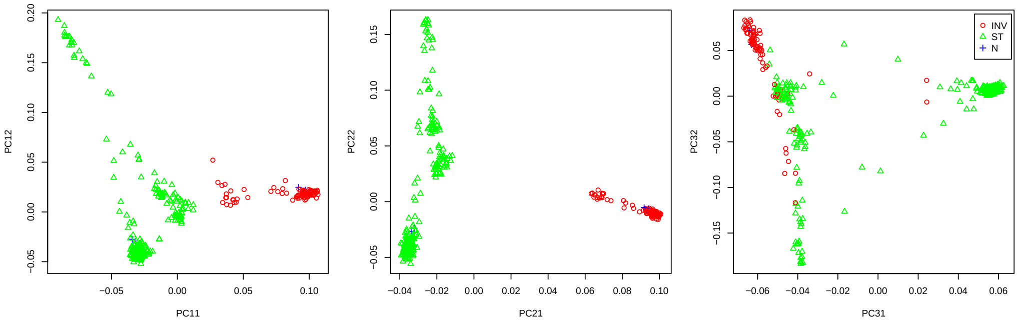

Figure 3: PCA plots colored by direction of In(2L)t direction for each peak in Chr2L MDS 2D plot. Red dots for samples with inversion, green dots for samples without inversion.

and 13154180bp. (Corbett-Detig and Hartl 2012) We recolored the PCA plots in Fig.2 a,b,c by the direction of the inversion for each sample using data from John Pool's lab. (Fig.3) *(citation?)* The results show that the clustering in PCA plots are mainly due to inversions for the red and green extremes in Fig.2a,b (especially for green extreme). What's more, we found that the two breakpoints of In(2L)t are located in the two clusters of the green points in Fig.1c. Similar results are found in other chromosome arms that have known inversions (Chr2R, Chr3L, Chr3R) in Drosophila. (See supplementary for more details) These facts show that the variation of population structure along the Drosophila genome is mainly due to inversions.

4.2 Human

We ran our method separately on all 22 autosomes. For chromosomes 3, 8, 15 and 17 have known inversions and the corresponding breakpoints (Antonacci et al. 2009). We found that the outlying windows in one-dimensional MDS plots coincide with the position of inversions for those 4 chromosomes.(Fig.4) As in Drosophila, the biggest source of variability in population structure along human chromosomes is inversion. Other chromosomes

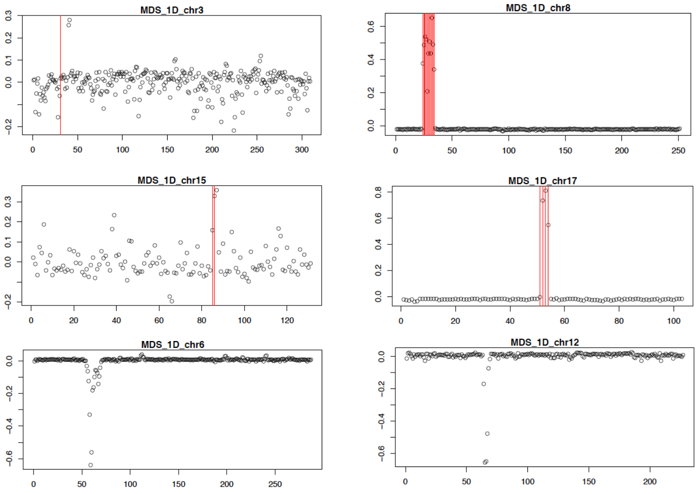

Figure 4: MDS results for human genome (chr3, chr8, chr15, chr17, chr6, chr12). The x-axis is the position on the genome in window number, and y-axis is the first coordinate of MDS. Black dots stand for windows, red lines show positions of inversions.

haven't had experimentally validated inversions, but there are many predicted inversions, and PCA might provide a way to validate those predicted inversion. (Ma and Amos 2012) (See supplementary for other chromosomes' MDS results for human)

Readers will wonder if the "triangle" looks the same in humans and Medicago as in Drosophila. Should we add them, or just put these in the supplement?

Do we know what happens when we run all 22 autosomes together?

4.3 *Medicago truncatula*

We checked the MDS results for all 8 chromosomes. *The results looked different than the other two species, with much less pronounced peaks...* We found the position of the peak for each MDS plot has a coincidence with the position of heterochromatic regions. This

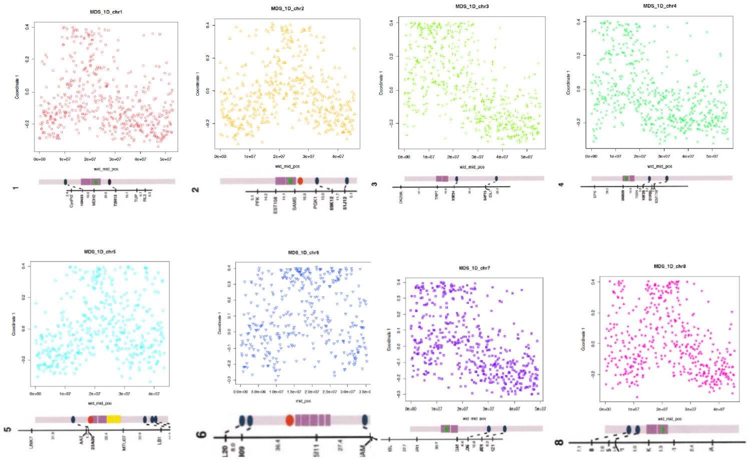

Figure 5: MDS results for the *Medicago* genome (chr1-8), using first coordinate of MDS against the middle position of each window along each chromosome. The bars under MDS plots are diagrams for each chromosome showing the relative position of heterochromatin (the deep purple part on each bar) (Kulikova et al. 2001, 2004).

means the population structure in the windows located in heterochromatin tends to have higher similarity, since those windows are closer in MDS plots. (Fig.5) Biologically, heterochromatic regions have lower gene density and may be less subject to selection (Kulikova et al. 2001; Paape et al. 2013). Then we checked the MDS results against gene density along the genome for each chromosome using gene models in Mt4.0 JBrowse. Numerically, the first MDS coordinate value is negatively correlated to the gene count for each window, which shows the correlation between MDS result and gene density. (Fig.6) This fact shows that in *Medicago*, the variation of population structure is correlated with heterochromatin and gene density, and is perhaps due to linked selection.

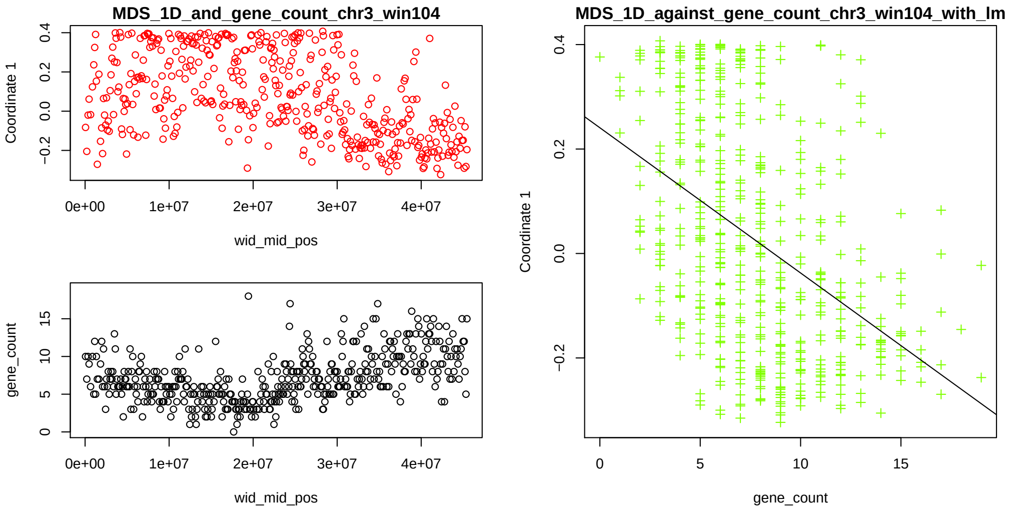

Figure 6: The first coordinate of MDS plotting against gene models gene count for chromosome 3 in *Medicago*. (See supplementary for other chromosomes' MDS results against gene count for Medicago)

5 Discussion

We've shown that local summaries of tree shape, i.e., kinship, vary significantly on a large scale along the genome, and that exploratory visualization of this variation can be useful.

There are many ways to visualize population structure; perhaps others would show different things.

Come back to some points in the Introduction.

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.

Nandakishore Kambhatla and Todd K Leen. Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516, 1997.

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.

Olga Kulikova, René Geurts, Monique Lamine, Dong-Jin Kim, Douglas R Cook, Jack Leunissen, Hans de Jong, Bruce A Roe, and Ton Bisseling. Satellite repeats in the functional centromere and pericentromeric heterochromatin of medicago truncatula. *Chromosoma*, 113(6):276–283, 2004.

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`.

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`.