Using Local PCA to Investigate Population Structure Along the Genome

Han Li, Peter Ralph

February 22, 2016

1 Introduction

About 37 years ago, Menozzi et al. (1978) first applied Principal component analysis (PCA) in population genetics to construct maps summarizing genetic variation (Menozzi et al. 1978). Nowadays, PCA is a widely used powerful non-parametric method to extract information from genetic data. PCA results are derived from the covariance matrix of genotype matrix, which describes population structure, demographic history, and pedigree tree and so on. The results of PCA can be directly related to the underlying genealogical history of the samples, such as coalescence time (time to most recent common ancestor) and migration rate between populations (Menozzi et al. 1978; Novembre and Stephens 2008; McVean 2009). Through dimension-reduction, PCA can identify key components of population structure, which describes how different samples are related, and plots of first two Principal components (PCs) can mimic the samples' geographic origin to some extent. The closeness of two samples in PC plots represents the similarity of their genetic information. Since population structure describes how different samples are related, and generally, samples living closer tends to be more genetically similar and thus tends to be clustered in PC plots (**??**).

Most time, PCA used in genome-wide association study (GWAS) is for stratification cor-

rection **?**. However, for different part of genome, they have different genetic features. First, each site of DNA may have different gene tree. For a DNA region, if individuals have SNPs closer in gene tree, they tends to be more close in PC projections for that region. Different DNA segments may have different gene tree for their SNPs and would result in 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. Third, if chromosome inversion exists, 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, thus results in different genome structure and 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. Therefore, we cut the chromosomes into windows with hundreds/thousands of SNPs and apply PCA locally on each window. 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 visualize the closeness for each window with regard to the population structure, we constructed matrix based on the first two PCs for each window and got the pairwise Euclidean distance between those matrixes, then use multidimensional scaling to reduces the pairwise distance matrix to lower dimension while preserve the distance information between windows as well as possible **?**. To interpret the results of MDS, we combine known genome feature information for each species, such as the distribution of inversions, heterochromatin chunk and gene density along the genome. For different species, we got different explanation for the variation of population structure along the genome.

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 Method

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

For human SNP data from POPRES **?**, 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). There are 3965 samples in total, (346 African-Americas; 73 Asians; 3187 Europeans; 359 Indian Asians); 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 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. 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 and recode it to matrices with 0,1,2 (and NA) for each of the 8 chromosomes.

3.2 Cut each genome into windows

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.

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

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

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 **?**. 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.

Table 1.

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.)

$$M_i = \frac{\lambda_{1i} PC1_i PC1_i^T + \lambda_{2i} PC2_i PC2_i^T}{\lambda_{1i} + \lambda_{2i}} \qquad M_j = \frac{\lambda_{1j} PC1_j PC1_j^T + \lambda_{2j} PC2_j PC2_j^T}{\lambda_{1j} + \lambda_{2j}} \quad (3)$$

The Euclidean distance between the constructed matrices for windows stands for the similarity of population structure for windows. Due to the property of eigenvectors,we could use the following method to calculate the pairwise distance greatly saving time and space.

$$V_1 = \sqrt{\frac{\lambda_{1i}}{\lambda_{1i} + \lambda_{2i}}} PC1_i \qquad V_2 = \sqrt{\frac{\lambda_{2i}}{\lambda_{1i} + \lambda_{2i}}} PC2_i \tag{4}$$

changed "Distance" to D

$$\begin{aligned} D_{ij} = \Big\{ & (V_1 \cdot V_1)^2 + (V_2 \cdot V_2)^2 + (U_1 \cdot U_1)^2 + (U_2 \cdot U_2)^2 \\ & - 2\left[(V_1 \cdot U_1)^2 + (V_1 \cdot U_2)^2 + (V_2 \cdot U_1)^2 + (V_2 \cdot U_2)^2\right] \Big\}^{1/2} \end{aligned} \tag{5}$$

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

Figure 1: The caption goes here.

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. (Fig1.a) 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 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.

Fig2.

There's a known inversion in Chr2L in *Drosophila*, In(2L)t, with breakpoints at 2225744bp and 13154180bp. **?** 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) 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.

Fig3.

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

Fig4.

4.3 *Medicago truncatula*

We checked the MDS results for all 8 chromosomes. We found the position of the peak for each MDS plot has a coincidence with the position of heterochromatic regions. This 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, hete-

rochromatic regions have lower gene density and may be less subject to selection **??**. 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.

Fig5.

Fig6.

5 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

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`.

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`.

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/`.

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/`.

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.

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.

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.

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.

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.