Local PCA finds heterogeneity in population structure along the genome

Han Li† and Peter Ralph†‡

† *Molecular & Computational Biology, University of Southern California*

‡ *Mathematics and Biology, University of Oregon*

http://biorxiv.org/content/early/2016/08/21/070615

https://github.com/petrelharp/local_pca

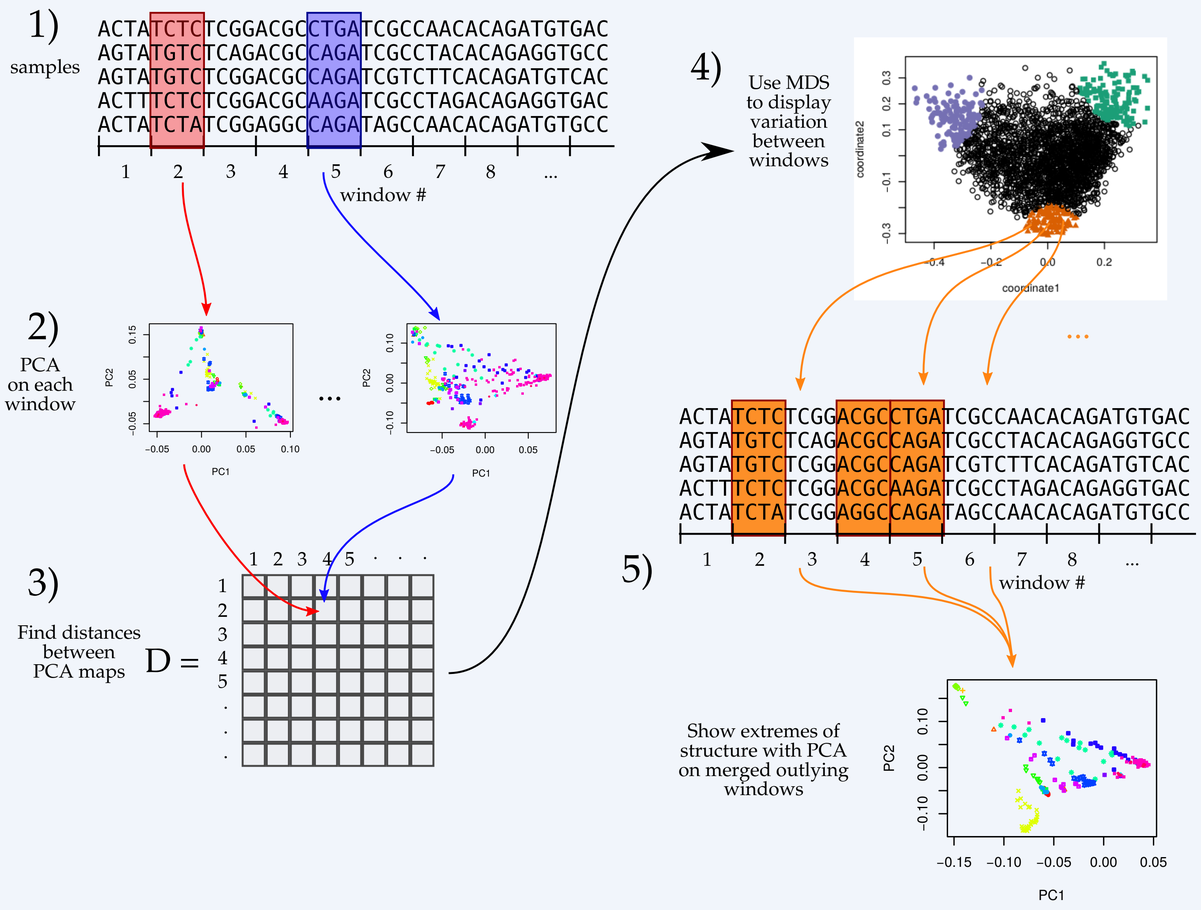
What is local population structure?

Population structure: "such matters as numbers, composition by age and sex, and state of subdivision" (Wright, 1949); but also commonly, the genetic patterns that result from this process. What aspects of demography should be included? Should natural selection? Incorporating differential survival or fecundity makes the concept less clear: does a randomly mating population consisting of two types that are partially reproductively isolated from each other have "population structure"?

Whatever the definition, selection causes the *effects* of population structure – *realized* patterns of genetic relatedness – to differ across the genome. Locally adapted alleles are selected against in migrants, increasing genetic differentiation between populations around this locus. Or, newly adaptive alleles spread first in local populations.

We aim to describe how patterns of mean relatedness vary along the genome. This works because: geography establishes similar patterns of relatedness across much of the genome, which are distorted locally by selection and other factors; we look for commonalities in these distortions (*not* outlier loci).

Our method



1. Code genome as $\{0, 1, 2\}$ and divide into windows (by # SNPs, bp, or cM).
2. Do PCA on each window: $\lambda_\ell^{(i)}$, $V_\ell^{(i)}$ the ℓ^{th} eigenvalue/vector of the covariance matrix for the i^{th} window.
3. Compute distances between windows:

$$D_{ij} = \| \sum_{\ell=1}^{k} \frac{\lambda_\ell^{(i)}}{\sum_{m=1}^{k} \lambda_m^{(i)}} V_\ell^{(i)} (V_\ell^{(i)})^T - \sum_{\ell=1}^{k} \frac{\lambda_\ell^{(j)}}{\sum_{m=1}^{k} \lambda_m^{(j)}} V_\ell^{(j)} (V_\ell^{(j)})^T \|$$

4. Visualize distances/similarities with MDS.
5. Highlight similar regions of the genome.
6. See how local PCA differs between regions by doing PCA after combining similar windows.

Notes:

1. Window size chosen by optimizing signal:noise
   ("signal:" variance in PC scores across a chromosome;
   "noise:" average jackknife-estimated standard error of PC scores by window);
   in applications, hundreds to thousands per chromosome.
2. Also used weighted PCA to remove the effects of oversampled groups without discarding data.
3. Results insensitive to rough choice of window size or unit of measurement.
4. Linear algebra reduced computation time of D from weeks to seconds.

Conclusion: Variation mostly caused by segregating inversions and linked selection; how much varies by species.

R package: available at `http://github.com/petrelharp/local_pca`

A.-S. Fiston-Lavier, N. D. Singh, M. Lipatov, and D. A. Petrov. *Drosophila melanogaster* recombination rate calculator. *Gene*, 463(1âĂŞ2): 18 – 20, 2010. ISSN 0378-1119. doi: http://dx.doi.org/10.1016/j.gene.2010.04.015. URL `http://www.sciencedirect.com/science/article/pii/S0378111910001769`.

J. B. Lack, C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool. The Drosophila genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, 199(4):1229–1241, 2015.

M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm, and E. H. Lai. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*, 83(3):347–358, Sept. 2008. doi: 10.1016/j.ajhg.2008.08.005. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556436/?tool=pubmed`.

H. Tang, V. Krishnakumar, S. Bidwell, B. Rosen, A. Chan, S. Zhou, L. Gentzbittel, K. L. Childs, M. Yandell, H. Gundlach, et al. An improved genome release (version mt4. 0) for the model legume *Medicago truncatula*. *BMC genomics*, 15(1):1, 2014.

S. 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`.

Results: *Drosophila melanogaster*

380 Pan-African samples from Drosophila Genome Nexus (Lack et al., 2015), `http://www.johnpool.net/genomes.html`.

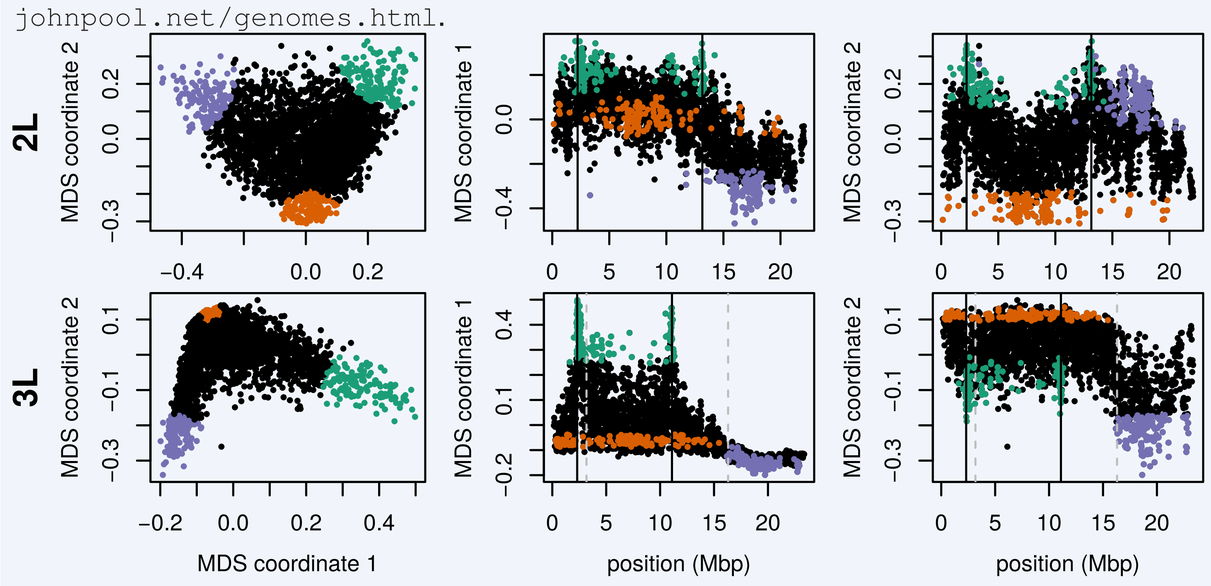


Fig 1: MDS, along the genome, for two chromosome arms. Vertical lines corrspond to inversions.

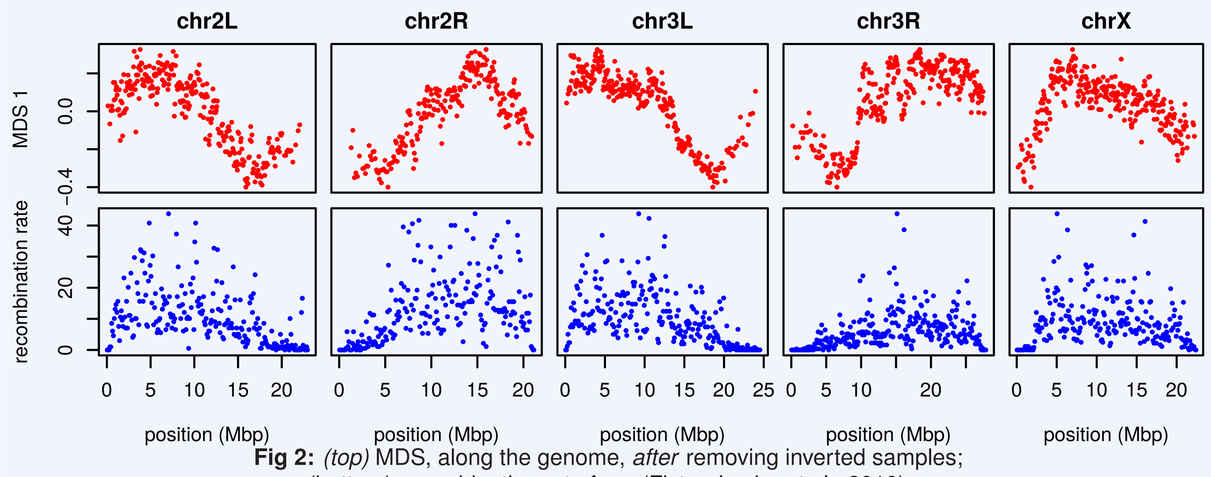


Fig 2: *(top)* MDS, along the genome, *after* removing inverted samples; *(bottom)* recombination rate from (Fiston-Lavier et al., 2010).

Results: *Homo sapiens*

The POPRES dataset (Nelson et al., 2008): 346 African-Americans, 73 Asians, 3,187 Europeans and 359 Indian Asians. (SNP array data; 500K markers)

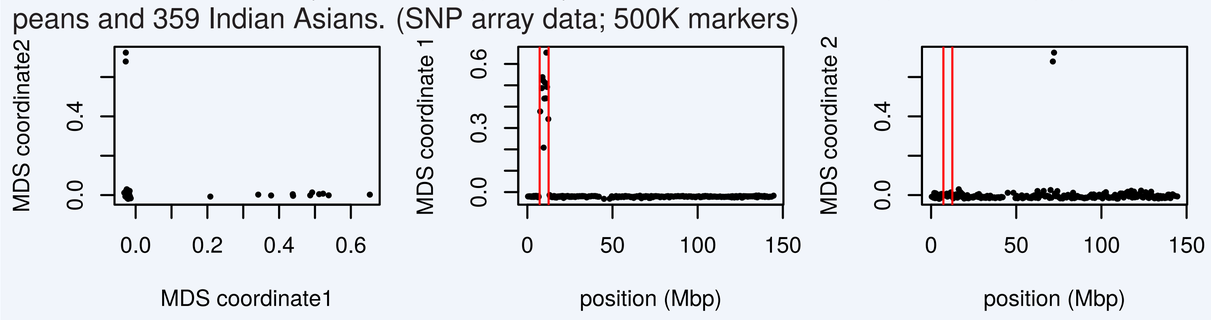


Fig 3: MDS on chromosome 8. Vertical line is a large segregating inversion.

Results: *Medicago truncatula*

263 samples from 24 circum-Mediterranean countries, from the *Medicago truncatula* Hapmap Project (Tang et al., 2014).

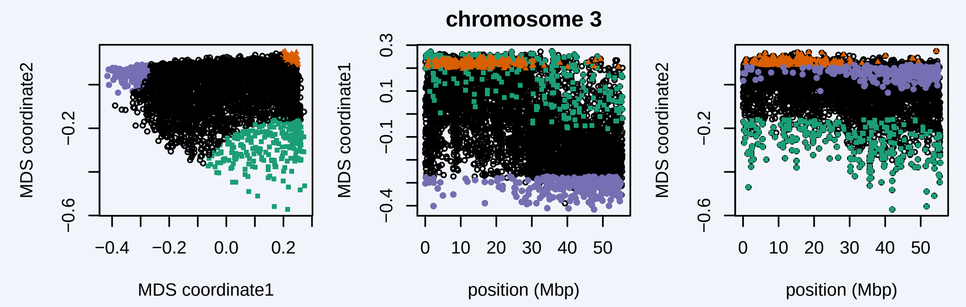


Fig 4: MDS along chromosome 3, showing continuous variation.

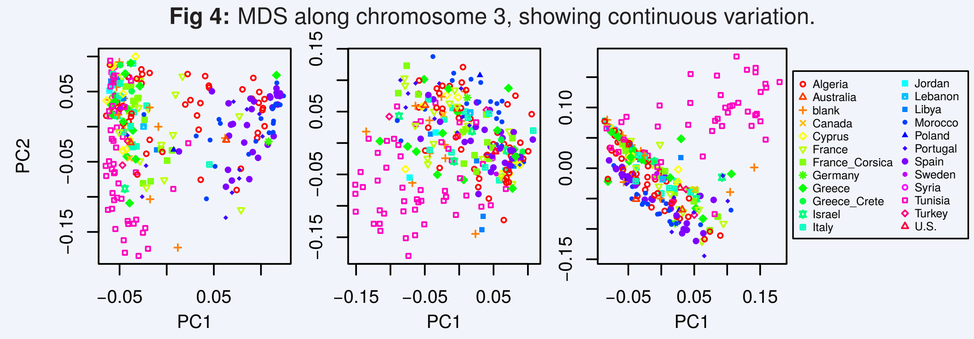


Fig 5: PCA plots corresponding to windows in the three corners of the MDS triangle.

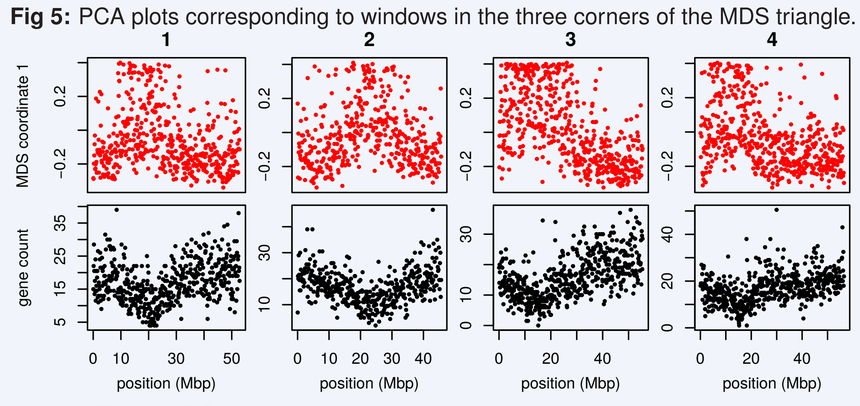


Fig 6: MDS score and local gene density along chromosomes 1–4.