Population Genetics Project #2

Wang et al. [2] studied patterns of genetic diversity in populations of Native Americans. Their sample included 422 individuals representing 24 different populations (Figure 1). They analyzed their data in the context of the HGDP-CEPH data set that I described in lecture [1] resulting in a analysis that included 1484 individuals.

You'll find a subset of the data set used in their analyses, consisting of a random sample of 100 microsatellite loci (out of the 684 included in the paper), on the course website in the format used by the program `STRUCTURE`: `wang-et-al.stru`. In addition, you'll find a file with the 1484 samples listed in the same order that indicates the group, area, and continent from which the sample was collected: `wang-et-al.pop`. There is a third file, `wang-et-al.dist`, that shows how the group numbers in the `STRUCTURE` file (the second column) relate to the continental areas in which that group is found. You will also find a ZIP file containing results from my analysis of these data with `STRUCTURE` for $K = 2, \ldots, 6$: `wang-et-al.zip`.

There are several levels of population structure in these data. Individuals are collected from *groups* (81), groups are located close to other groups in *areas* (37), and areas are found in large areas on several *continents* (7). Using the data provided here and what you learn from reading the two papers associated with this project, answer the following questions:

1. How much inbreeding is there within groups?

2. How much of the total diversity present is found within the groups that were sampled?

3. What level of the hierarchy shows the greatest amount of differentiation: continent, area, or group?

4. How many genetic clusters is it worthwhile to distinguish in these data?

5. Are there genetic clusters found in more than one continental area?

6. Use `dapc()` in `adegenet` to perform a Discriminant Analysis of Principal Components and compare the results to what you found with `STRUCTURE`.

7. What does the pattern of genetic similarity among within and among continental areas suggest about either how frequently individuals move among these areas or how recently

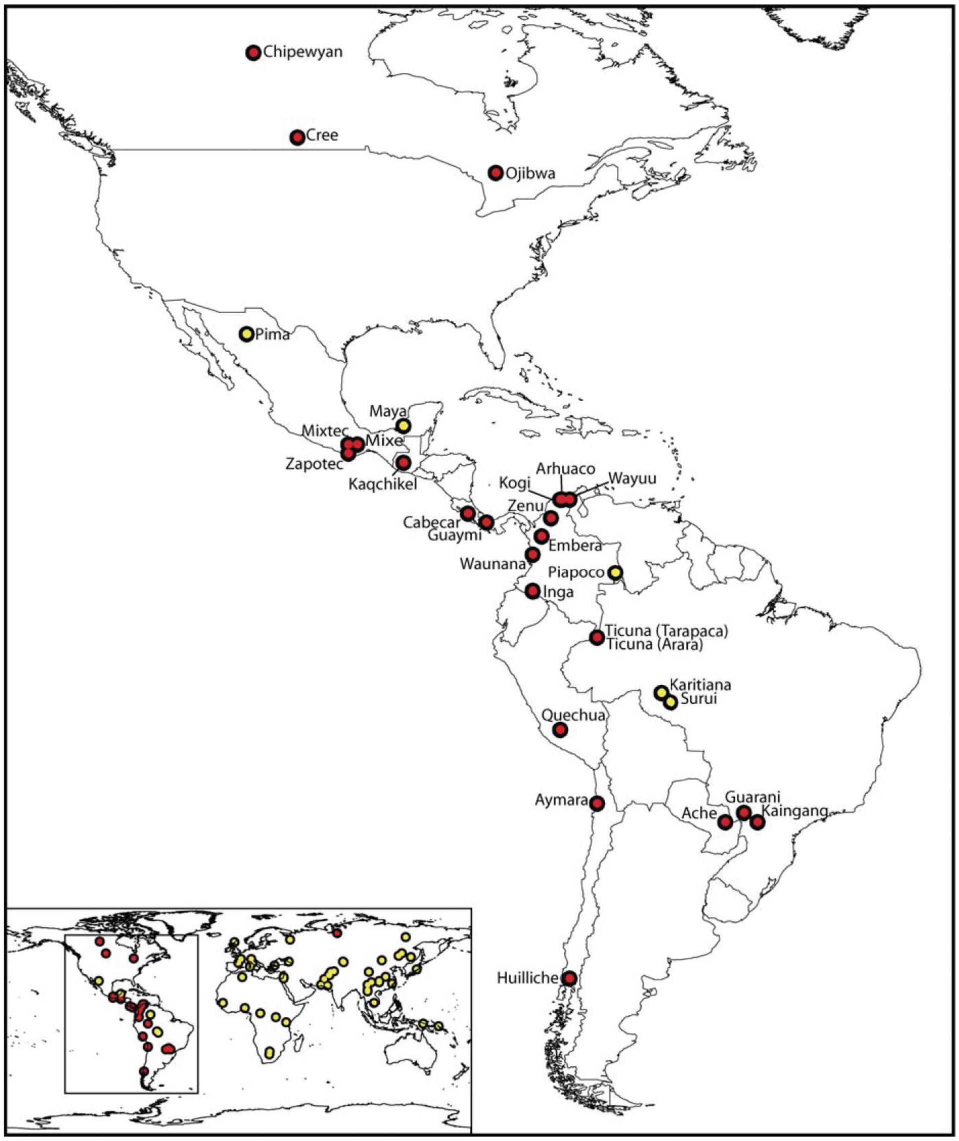

Figure 1: The geographic distribution of the 78 populations included in the study from which data for this project were derived. The 29 populations from the Americas that were new to the study are shown in detail.

populations in these areas became isolated from one another? What insight does this provide into how Native Americans arrived in the Western Hemisphere?

Hints

- Use the R script (analysis.R) to read data from wang-et-al.stru and prepare it for analysis with hierfstat.

- Visit http://clumpak.tau.ac.il/index.html and upload the wang-et-al.zip where it asks for a ZIP file and wang-et-al.dist where it asks for a "labels file for DISTRUCT (optional)." Enter an e-mail address, hit submit form, and you'll get an e-mail after awhile with links to results of the analysis, a PDF file showing the clustering for each K detectedModesSummery.log.[1] Examine LnProb mean for each K to help you identify the number of clusters suggested by these data.[2]

- To mimic the classification of individuals in STRUCTURE using dapc(),

  1. Determine the number of groups and store the result:

     ```
     wang_grp <- find.clusters(wang_fst, max.n.clust = 20)
     ```

  2. Chant a little bit of magic:

     ```
     pop(wang) <- NULL.
     ```

  3. Run the analysis:

     ```
     wang_dapc <- dapc(wang_fst, wang_grp$grp)
     ```

  4. Visualize the result:

     ```
     scatter(wang_dapc, pch = 20, cell = 0, cstar = 0, clab = 0,
     leg = TRUE, posi.leg = "topleft", posi.da = "none",
     grp = other(wang_fst)$continent)
     ```

[1] Yes, it's spelled "Summery," not "Summary."

[2] Don't worry if the number of clusters suggested by this analysis differs from what is suggested in the paper. These results are based on less than one-sixth the number of loci included in the analysis for the paper.

- We won't discuss how the amount of population differentiation depends on gene flow or common ancestry for a couple of weeks, so we're not expecting a great amount of detail. I'm sure you already know that the more gene flow there is between populations, the more recently they became isolated from one another, or both, the more genetically similar they will be with one another. Use that intuition to answer the last question?

References

[1] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

[2] Sijia Wang, Jr. Lewis, Cecil M., Mattias Jakobsson, Sohini Ramachandran, Nicolas Ray, Gabriel Bedoya, Winston Rojas, Maria V. Parra, Julio A. Molina, Carla Gallo, Guido Mazzotti, Giovanni Poletti, Kim Hill, Ana M. Hurtado, Damian Labuda, William Klitz, Ramiro Barrantes, Maria Ctira Bortolini, Francisco M. Salzano, Maria Luiza Petzl-Erler, Luiza T. Tsuneto, Elena Llop, Francisco Rothhammer, Laurent Excoffier, Marcus W. Feldman, Noah A. Rosenberg, and Andrs Ruiz-Linares. Genetic variation and population structure in native americans. *PLOS Genetics*, 3(11):e185, 2007.

Creative Commons License