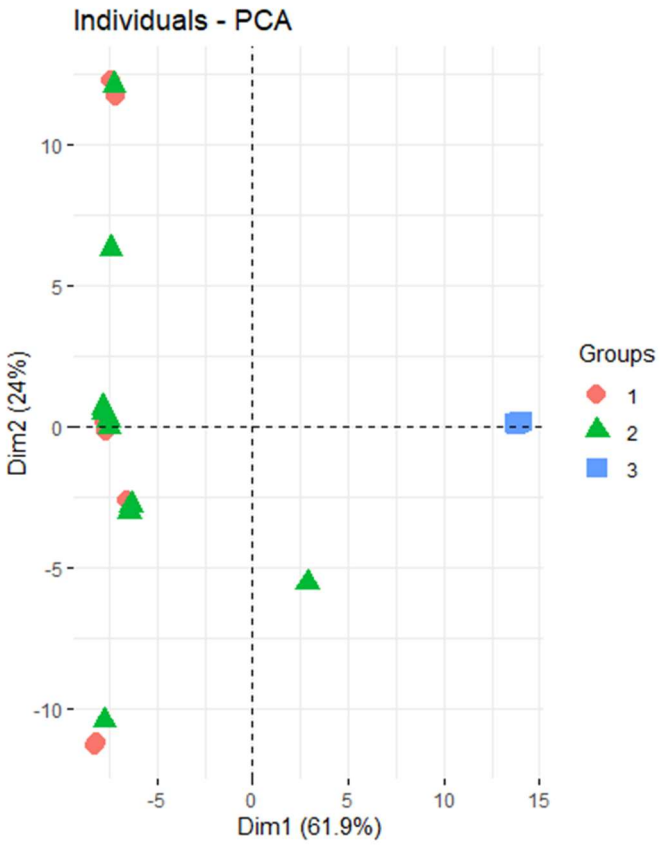
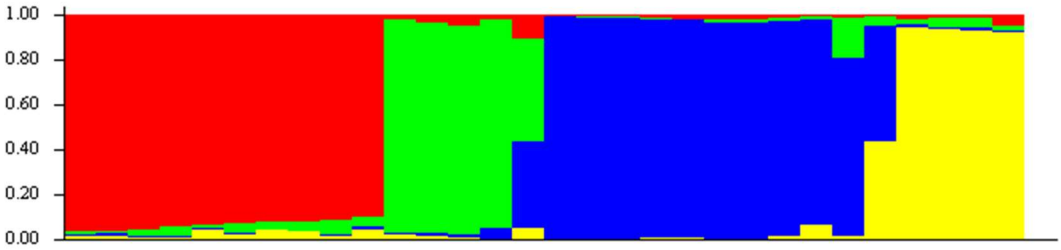
three_pop_migration:

Structure identified k=4 and k=5 as the most likely number of populations. Judging from the name, this is probably a flawed estimate. Furthermore, based on the runs on mystery files, Structure may have continued for a while to claim that k=6 or 7 were more likely. The plot created using PCA suggests three true populations with a lot of migration between two groups and a little migration between another pair; no migration occurred between the third pair, in appearance. The script itself does seem to agree with this claim.



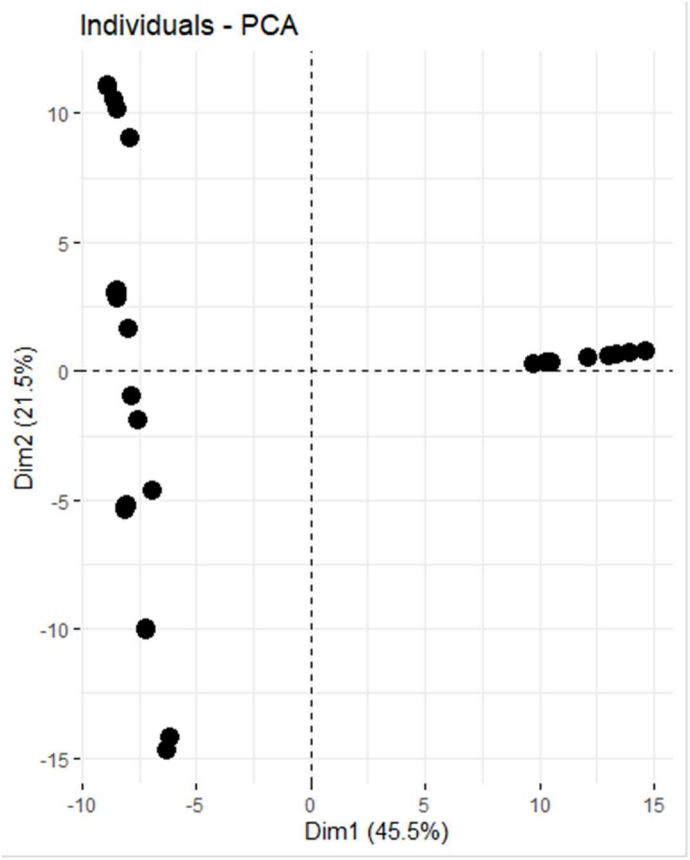


The bar graph for k=4 is given above, and the four populations in question probably correspond to the three true populations as well as the intermediate genetic group caused by the higher migration rates between two in particular. Because of the lack of red bars whose magnitude is between 0.2 and 0.8, it seems reasonable that the red blocks correspond roughly to group 3 in the PCA above. The blue blocks

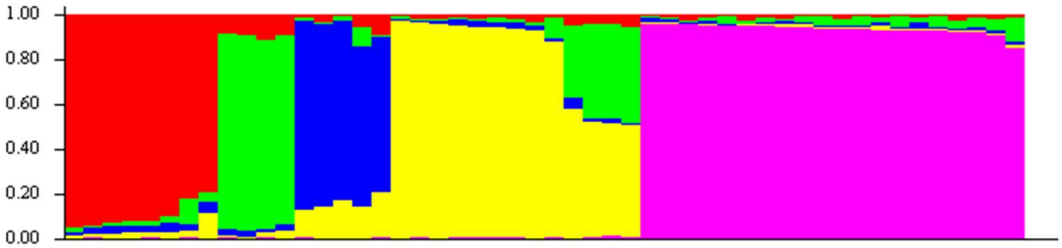
probably correspond roughly to the intermediate group described by the PCA, while the green and yellow blocks probably correspond to the two groups sharing high migration rates.

Write-your-own:

I wrote a script involving 5 subpopulations, named 1, 2, 3, 4, and 5 for convenience, where 1 and 2 each had migration rates between them of 0.005, 2 and 3 had 0.005 between them, 4 migrated to 5 at 0.005, and 5 migrated to 4 at 0.001. Each subpopulation has a consistent size of 500 individuals. Unfortunately (or, alternatively, naturally), populations 1, 2, and 3 seem to be colinear in the graphical representation of PCA:



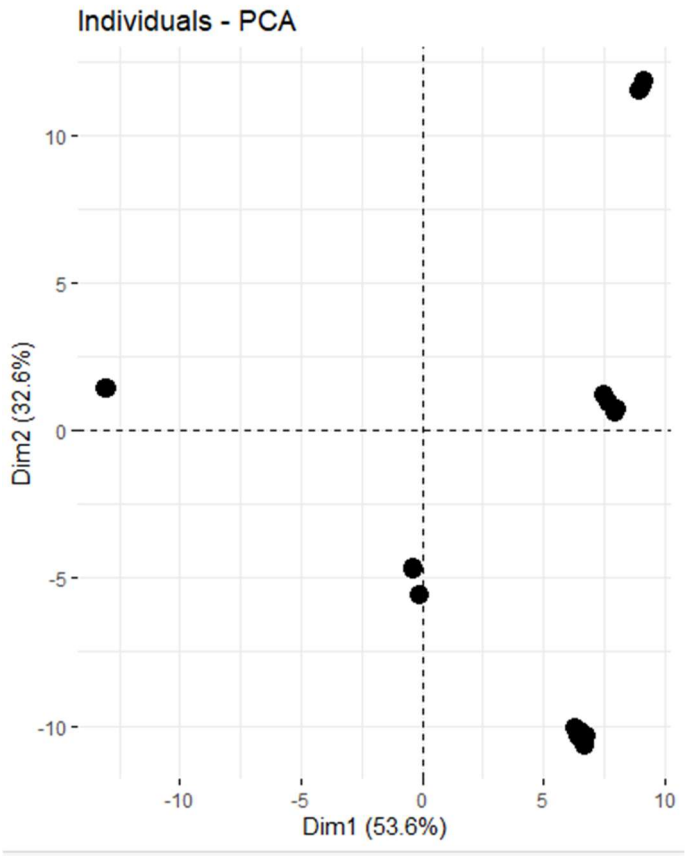
This plot alone does not seem to be clearly representative of the simulation in question, probably because 3 of the populations are colinear. It does seem to suggest, however, that the direction of migration is difficult to ascertain using PCA alone. Analysis through Structure is, perhaps, more enlightening. The smallest p-values seem to range between k=4 and k=6, bottoming out at k=5, which, interestingly, is representative of the population in the absence of migration. The bar graph for the k=5 run is given below:

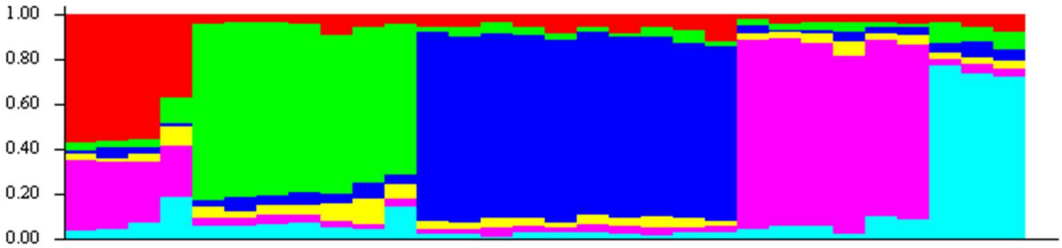


The fact that all 5 colors appear in all five regions immediately suggests that the regions do not correspond perfectly to the populations, and the vastly different sizes of blocks of color likely indicate that regions do not correspond to populations well, as it is not reasonable to conclude in this simulation that one population has contributed this much more to the gene pool than any of the others. An interesting part of this bar graph is the relatively large portion which is about half green and half yellow. It would be more interesting if the rest of the graph indicated some sort of significance.

mystery1: Structure identified k=6 or 7 has the highest, with a definite upward trend

The first run of Structure used 2 through 5, inclusive, for k, but the p-value increased steadily over this domain, so I decided to run it again through k=7. P-values continued to increase through k=7, but PCA seemed to indicate an upper bound of k=6 as a possibility (despite k=7's p-value's being lower than that of k=6), so I stopped there. The PCA plot seems to indicate three truly distinct subpopulations, with no migration occurring within exactly one of the pairs. Since migration in both directions within each of the remaining 2 pairs would result in some mixing within all three pairs, it is reasonable to conclude this is not occurring. Therefore, individuals from the lower right population probably migrate to the other two populations, but none return, as this would pull the intermediate groups toward the middle.

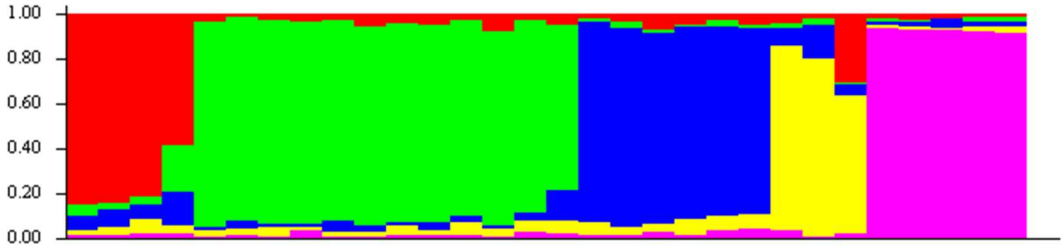




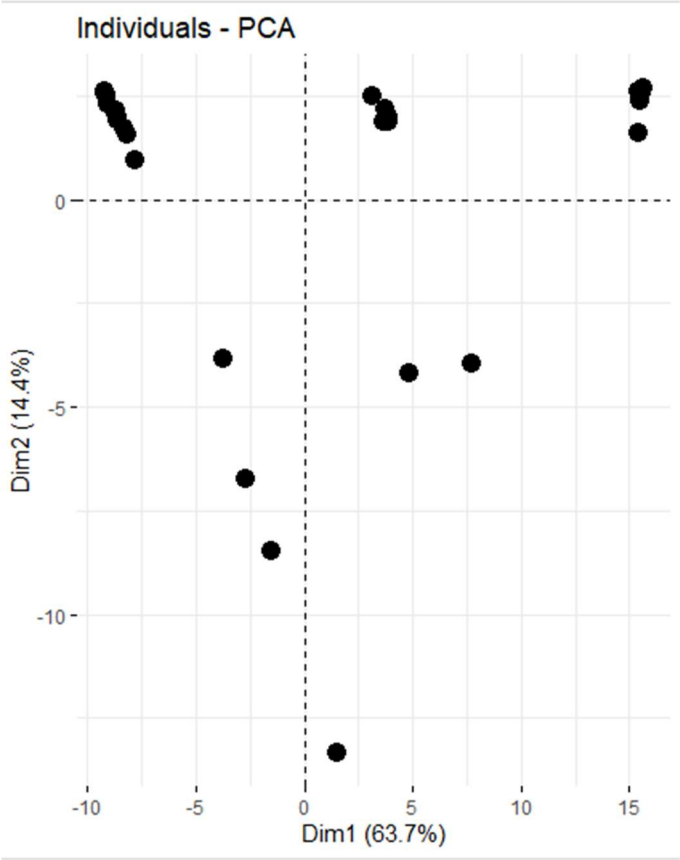
The Structure bar graph, shown here for k=6, seems to indicate more mixing than the PCA plot above does; this is likely the result of overestimation of k. The PCA plot given above indicates a maximum of 5 genetically differentiable groups, so populations described by this bar graph probably do not correspond to groups with any true identity.

mystery2:

Using k=2 through 10, Structure's simulation claims that the p-value peaks (or whatever the opposite of that is) between k=6 and k=8. The bar graph describing the run assuming k=6 appears (sorted by Q) appears to show… 4 distinct groups. In contrast, k=5's run displays 5 differentiable groups, which seems slightly less insane. This bar graph is displayed below:

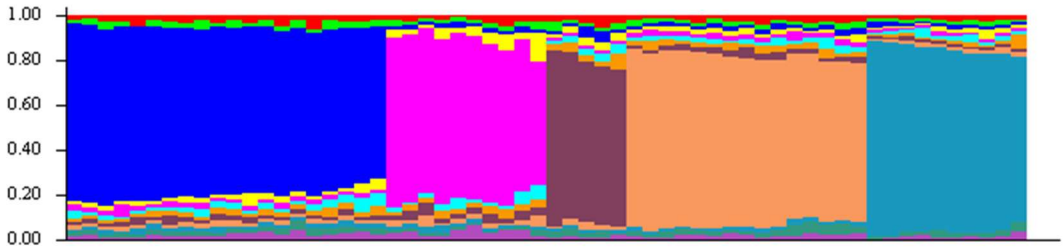


Pink seems to be associated with a group that is more independent than the others, possibly indicating that it is less involved in migration overall. The same can be said for green, although to a lesser degree. The opposite is true for red, blue, and, especially, yellow. These groups seem to be experiencing larger degrees of migration. Observing the plot given by PCA below, the most isolated seems to be the group in the upper left corner. The second-most is in the upper left corner; it may seem reasonable, then, to conclude that these two groups probably correspond pink and green, respectively. However, the bar graph above does not show any significant overlap between green and pink, and the PCA plot shows a group intermediate between the two corner groups. The yellow bars here are perhaps the most interesting, as they reach (barely, if that) only the .80 mark at the maximum. This could indicate that groups that perhaps should be considered distinct are being represented by fewer than actually exist. A better representation of this data may be given by the graph corresponding to the k=3 run, as genetically distinct groups would be isolated from one another, but their mixing may be more straightforwardly observed as bars of mostly two colors, rather than messy blobs like the yellow above.



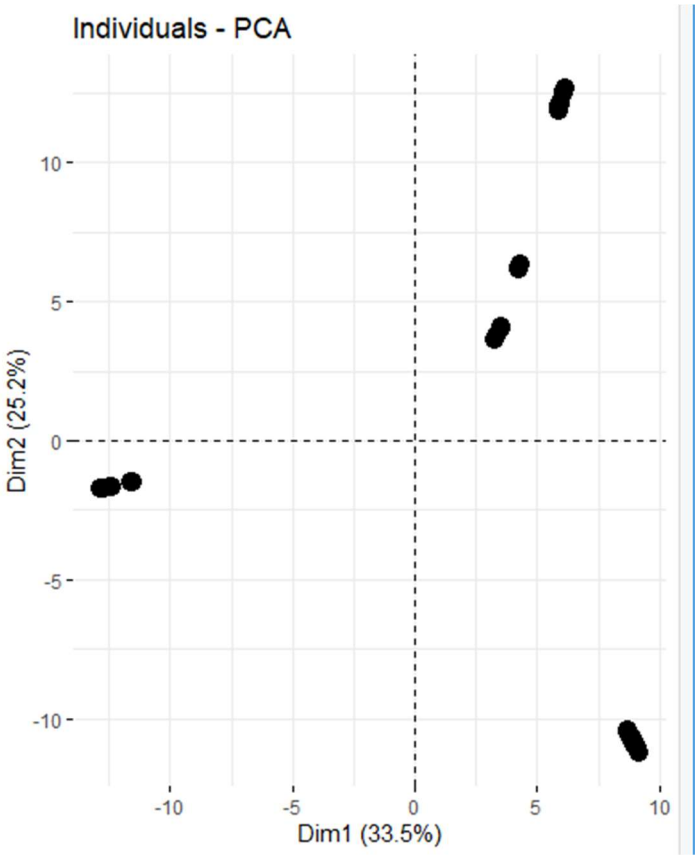
mystery3:

This data set has been the most entertaining. The runs from k=8 to k=12 have the lowest p-values, with surprisingly little variation, and an extremely fast dropoff afterward. Observation of any one of the corresponding bar graphs, however, claims “the numbers are lying. There are five groups, no more, no less. Just look at this beautiful 5-colored rainbow.” Partially for entertainment purposes, I’ve included the bar graph corresponding to the k=12 simulation below:



This graph clearly depicts 5 distinct groups. The graphs of the k=13 and k=14 cases look disturbingly like illusions of straight lines intended to look crooked, even when sorted by the same parameter, which probably has something to do with the super high p-values. This probably indicates that there are at most 12 groups of individuals that cannot be separated into smaller groups of individuals significantly more similar within the group than between groups. This in turn indicates that the population in question consists of far fewer groups in reality.

The script provided, using PCA, provides the following output:



This seems to indicate three extreme regions and one or two intermediate regions. The intermediate regions do not appear to be halfway between any of the three extreme regions; rather, they seem to be about twice as close to one region as they are to either of the other two. It seems as though there may be migration to the region in the upper right corner from the other two regions.