Depicting simultaneously similarity, diversity, ancestry, and admixture?

Peter J. Taylor
peter.taylor@umb.edu

Follow this and additional works at: https://scholarworks.umb.edu/cct_sicw
Part of the Biological and Physical Anthropology Commons, and the Molecular Genetics Commons

Recommended Citation
https://scholarworks.umb.edu/cct_sicw/16

This Article is brought to you for free and open access by the Critical and Creative Thinking Program at ScholarWorks at UMass Boston. It has been accepted for inclusion in Working Papers on Science in a Changing World by an authorized administrator of ScholarWorks at UMass Boston. For more information, please contact library.uasc@umb.edu.
Depictions of human genetic relationships
Quick explorations that invite further graphic, conceptual, and interpretive work

PETER J. TAYLOR
Depictions of human genetic relationships:
Quick explorations that invite further graphic, conceptual, and interpretive work

Peter J. Taylor
Graduate track in Science in a Changing World
University of Massachusetts, Boston, MA 02125, USA.
peter.taylor@umb.edu

April 2011

Abstract
Can any depiction of genetic relationships among humans allow simultaneously for similarity, diversity, ancestry, and admixture (i.e., groups that had split mixing again)? I asked this question while considering the depiction by Tishkoff and her collaborators of the branching pattern of human groups within and out of Africa—mostly within Africa (Campbell and Tishkoff 2010). This think-piece presents my quick explorations of alternative depictions of human genetic variation. I offer it not as a polished investigation, but as an invitation to others to undertake further graphic, conceptual, and interpretive work. As will be evident by the end, the work needed to create and to understand alternative depictions of human ancestry is important because the current graphic conventions privilege a racialized view of human diversity. (Note: Some references have been updated since the original 2011 text.)
The starting point

Campbell and Tishkoff (2010) present the following diagram that, using data about genetic variation among present-day humans, reconstructs their ancestry in and out of Africa.

![Diagram of human ancestry](attachment:image)

**Figure 1.** The Recent African Origin Model of modern humans and population substructure in Africa (with addition by author of letters to label the groups), from Campbell and Tishkoff (2010)

The most obvious message is that there are many more branches leading from the ancestral within-Africa human population to current African groups than there are to groups in the rest of the world, that is, to groups derived from people who migrated out of Africa at some point after 100,000 years ago. This branching pattern suggests that, if we were to divide human genetic diversity into a small number of groups each having a similar amount of genetic diversity, say, five groups, then most of these groups would
be African. (Indeed, four would be from Africa and the fifth would be a combination of an East African group and the non-African groups.) This finding seems to discredit the genetic reality of traditional races, where “traditional” has varied greatly but, in general, separates Africans, Asian, Australian, and Europeans or “Caucasians.”

Different messages can also, however, be drawn from the same diagram. The non-Africans are placed out on the right, taking up 40% of the horizontal scale, colored lighter, and with very few of the horizontal cross-linking bars (which denote gene-flow between populations) connecting them to the African groups. All these features suggest that the African groups can still be lumped together. Someone who promotes a traditional racial classification (or someone from the last two centuries who promoted such a classification) might not be at all troubled on learning that their African race is really a collection of 13-14 races. All they need is the idea that something special genetically happened in the branch that left Africa; the groups left in Africa are united in lacking that something. It is these Africa versus non-Africa messages that are problematized by the explorations that follow.

**Exploration 1: Rearranging the horizontal sequence of a tree diagram**

The Campbell-Tishkoff diagram of human ancestry branches out like an upside-down tree from a common ancestral group into 18 groups today. In Figure 2 we see the tree for the first three forks, where AR is an abbreviation for a group that includes all the ancestors of groups A through R; NR for all the ancestors of groups N through R; etc.

![Figure 2](image-url)
Now, the branches at any fork can be flipped so Figure 3 conveys the same information about ancestry and branching.

![Figure 3](image)

Notice that the second variant does not convey the impression that the branch that is ancestral to the non-Africans, i.e., NR, is more different from the branches ancestral to the African groups, i.e., AB, CC, DM, than these branches are from each other. Although the lineage that ended up at CC (the ancestor of group C) branched off earlier than the lineage leading to NR, there is nothing in the ancestry diagram that says it should be more similar genetically to AB than to NR.

![Figure 4](image)

If we exclude diagrams with crossing over of branches, such as Figure 4, there are four distinct reorderings of the four branches that preserve the sequence of the branchings.
There are $2^{16}$ or 65536 reorderings of the full set of the 18 current groups. The point is not that we need to find one correct ordering from among such a large set. The lesson is that no lessons should be drawn from the order along the bottom of a branching diagram that is not already contained in the sequence of branches above. (This last caveat ensures that diagrams with crossing over are excluded because they suggest that the two branches at a fork are further away from each other than to one of the earlier branches, which goes against the information contained in the sequence of branches.) It is not easy, however, to convince one’s brain not to give significance to these horizontal positions. This cognitive weakness gives rise to the next explorations.

Exploration 2: Arranging the groups on the ancestry tree so that distance reflects (to some extent) the time since branching.

In order to allow for the 65638 reorderings of the 18 groups at the base of the Campbell-Tishkoff diagram, we might think of a mobile with each pair of branches able to revolve around the position of its most recent common ancestor, which will itself be moving as it revolves with another branch around its common ancestor. Virtual mobiles can be built using a National Gallery of Art (2019) website. The software allowed me to replicate the Campbell-Tishkoff diagram down to the level of 4 branches, i.e., AB, CC, DM, and NR, making the distance between any pair of branches, i.e., (DM,NR), (CC,DR), (AB,CR) proportional to the time since they diverged. The mobile software allows one to view the mobile from above as well as from the side. In the following
Figure 5. Snapshots from above of the mobile down to the level of 4 branches (redrawn into black and white). The open ball is AB; the striped one is CC; and the other two are DM and NR. The black dots are the pivot points for the arm of the mobile.

Making the distance between any pair of branches proportional to the time since they diverged differs from the original Tishkoff diagram, which has, for example, A, B, and C close together at the bottom even though the common ancestor of A and B (i.e., AB), branched off 150,000 years ago from the ancestor of C (i.e., CC)—plenty of time for genetic divergence to have occurred. The distance relationship between members of a pair does not mean that the distance between every pair of the four groups at the base of the mobile is equal to the time since their common ancestor. That property is not possible to achieve in a tree depiction. Instead, the mobile simply serves to remind us, as in Exploration 1, that no lessons should be drawn from the order along the bottom of a branching diagram that is not already contained in the sequence of branches above.
Exploration 3: Arranging the groups on the ancestry tree so that distance reflects—to some extent—the time since branching.

Figure 6 uses the same distance relationship as for the mobile but now it holds for all pairs of branches in the full tree. (Of course, a more refined analysis might allow for different speeds of divergence from the common ancestor down different branches. This could also be depicted in this same form.)

Figure 6.
As in the previous exploration, the distance relationship between members of a pair does not mean that the distance between every pair of the 18 groups at the base of the mobile—there are 153 such pairings—is equal to the time since their common ancestor. The crossing over makes that limitation obvious. It is easy to see that, for example, the closeness of H and L is not because they share a recent common ancestor.

**Exploration 4: Arranging the groups on the ancestry tree in two dimensions so that distance reflects the time since branching (somewhat).**

By spreading out the tree into a fan, the distance between every pair of the 18 groups at the base of the diagram can be made much closer to the time since their common ancestor.
Figure 7.

The two-dimensional branching eliminates the crossing over that made the tree in Exploration 3 difficult to read. However, the depiction is far from perfect. For example, C
and N end up close even though their common ancestor was almost as distant in time as could be. A sophisticated algorithm might arrive at a better fan than the first attempt above, but we could never get around the fact that, at each branching point, the branches could be flipped (e.g., E or F could be made close to D instead of G).

**Exploration 5: Superimposing genetic variation on the ancestry diagram**

When compared with the Campbell-Tishkoff tree of human ancestry, the two-dimensional branching depiction of Figure 7 greatly improves the degree to which the distance between groups is proportional to the time since the groups shared a common ancestor, but cannot eliminate spurious appearances of similarity. Even if that limitation were put aside, it is important to note that *the two-dimensional branching depiction omits the genetic variation around the midpoint of any branch*. Such variation is hinted at in the Campbell-Tishkoff diagram in two ways: 1. The relative thickness of the branches—the thick trunk at the top indicates more genetic variation in the ancestral group than the think tips in the branches at the base; and 2. The density of the color of the branches—the deeper blue (or black in this publication) indicates more genetic variation than a lighter-shaded branch. (The migration out of Africa involved a small group that brought with it only a small subset of the genetic variation in the African ancestral branch from which it broke off.)

Although variation around the group’s midpoint is suggested by the preceding two features, the Campbell-Tishkoff ancestry diagram does not in any way convey the fact that, on average, for any genetic locus roughly 5/6 of the variation is within a population, 1/12 is within a region, and only 1/12 occurs among regions (using figures dating back to Lewontin 1972, but affirmed by subsequent work, such as Hofer et al. 2008). To convey this variation, we could add “aprons” around the midpoints of the branches in the two-dimensional depiction. In Figure 8 aprons are added, but around groups A and H only.
Figure 8.
Even without aprons around every terminal branch, a key point is evident, namely, ancestry trees show the genetic midpoints of branches and thus mostly hide the large amount of genetic variation not captured by the branching pattern. (It is because of this point that I have not explored whether the available data would allow us to calibrate the apron size to match the different degrees of genetic variation within the groups at the ends of the branches.) As a corollary of this point, such variation makes it difficult, on the basis of a random selection of genetic loci, to assign an individual to one branch or the other. We need to say difficult, not impossible; it is subject to more errors than to correct assignments. We also need to say random selection of loci because clearly there must be some genetic differences that are specific to a branch; otherwise we would not be able to trace ancestry patterns at all. To spell this out: If there are mutations that are very common in some people and rare in others, a tree can be made that captures the most likely branching pattern, which is one that assumes the least reversions (i.e, mutation in one direction, mutation back again to the original condition). We can make such trees even if most genes vary in ways that bear no sign of that branching.

**Exploration 6: Superimposing changes in genetic location on an ancestry diagram, a simulation**

The two-dimensional branching depiction in Figure 7 is far from perfect and the size of the aprons in Figure 8 was determined by a back-of-the-envelope method. Nevertheless, the combination of the two-dimensional branching and aprons holds some promise for allowing simultaneously for similarity, diversity, and ancestry, which was part of the original question motivating this series of explorations. To realize this promise we have to revisit the issue of branches crossing over.

In a branching diagram, we sought to minimize the crossing over of branches because the resulting diagram contradicts the information contained in the sequence of branches by suggesting that the two branches at a fork are further away from each other than to one of the earlier branches. However, the inclusion of aprons that overlap allows us to
embrace branches that cross over. Consider the following model, which also allows for evolution along branches to happen at different rates:

In a branching process, each group breaks into two. Imagine that the new groups are small so that by genetic drift, that is, by chance, all members end up with the same variant at some genetic locus (i.e., position on the genome), that is, this locus does not contribute genetic variation. The population eventually grows larger and genetic drift ceases to be significant. Each of the new groups then represents a subset of the variation existing in their common ancestor group. If we discount new mutations for now, none of the branched-off groups can have more genetic variation than groups from which they are descended. (A recent model of language evolution mirrors this model in many respects; Atkinson 2011.) Random decisions can be used to generate directions of branching in two dimensions and the genetic distances of each branch from its most recent common ancestor. Figure 9 is one simulation of the model.

Figure 9.
No aprons are drawn around the midpoints of the groups, but let me note that the variation of the original population spans a space five times as wide as the area shown in the diagram. (A slide show at http://wp.me/pPWGi-m0 builds up the messy web branching from one group—AR—to two—AB and NR—and so on, step by step.) Of course, this is only a simulation. The actual genetic data might yield a two-dimensional web that is quite different.

**Exploration 7: Superimposing genetic variation on the ancestry diagram from a simulation**

Figure 10 comes from the same random simulation used to generate Figure 9. This time aprons are drawn around the midpoints of the groups A to R at the bottom of the ancestry tree (but not around their ancestors) so that Figure 10 provides a picture of similarity, diversity, and ancestry.
As in Figure 9 the two dimensions (horizontal and vertical) stand for the genetic variation of the whole set of populations. The variation of the original population (which would extend about 20% past the largest circle) is reduced after the branchings that have brought us to the present, but there is still great overlap between most groups. In particular, the descendants A and B of the group AB, which branched off early, show variation that subsumes that in the rest of the groups. A careful viewer might notice that
there are some circles that do not overlap at all, as if to say these groups share no genetic variation. This is an artifact of an adjustment made to the simulation that reduced the variation remaining after each branching in order to ensure that not all the circles would extend beyond the web. This adjustment, made with graphic reasons in mind, increased the ratio of between-groups-differences to average-within-group variation well beyond what we find in the actual human world.

**Exploration 8: Undesirable messages in the conventional ancestral tree of human groups**

The original Campbell-Tishkoff diagram of human ancestry is certainly easier to read than the reticulating web of Figure 9, let alone the web overlaid with aprons in Figure 10. We could try to improve the readability of these figures by helping audiences to become familiar with the graphic conventions and by using technology like the slide show to display the branching and replacement of ancestral aprons with those of their descendants. Indeed, I believe that it is important for all to work on being able to read reticulating webs of human ancestry and genetic variation because of an undesirable message built into the simpler branching diagram.

To expose the undesirable message, consider the horizontal links in the Campbell-Tishkoff diagram, which represent admixture or gene flow between branches. The branching pattern can be extracted from the genetic data only because these flows are not so large as to obscure the genetic mutations or other differences that arose over time after each branching. Indeed, to ensure that this is the case, some studies of human genetic variation involve data from the special subset of people who live in the same place as, say, all their great-grandparents. The reticulating web with aprons likewise relies on a branching pattern that can be discerned despite the potentially confounding effects of gene flow. Yet the aprons remind us of variation around the midpoint of each group—variation that may well have been enlarged by gene flow.
Now, there are some branching patterns that are subject to minimal or no gene flow, namely branching of species or higher taxa from ancestral taxa. We are all familiar with such evolutionary trees, whether for the classes of vertebrates or the range of liverwort species (Figures 11 and 12).

Figure 11. Phylogeny of vertebrates (original source unknown).
Our familiarity with these trees invites us to think—perhaps subconsciously—about human genetic ancestry as if the branches are like separate species. There is a long history of scientific arguments that human races are separate species, or that the branches of the human tree achieved human status at different rates. As Desmond and Moore (2009) have shown in *Darwin’s Sacred Cause: How a hatred of slavery shaped Darwin’s views on human evolution*, the debate was especially heated during Darwin’s adult life. Darwin’s view of descent from a single common ancestor was a minority view, discredited to some extent by its association with literal interpretation of the bible’s account of Adam and Eve, but more so by its association with anti-slavery movements.

The multiple origins debate continued, shifting in form, into the 20th century. Carleton Coon, a physical anthropologist who died in 1981 after a long career as a professor at Harvard and University of Pennsylvania, wrote in 1962 that *Homo erectus* evolved into *Homo sapiens* five separate times "as each subspecies, living in its own territory, passed a critical threshold from a more brutal to a more sapient state." The multiregional hypothesis is a more recent variant.

Ideas about multiple origins for humans are not the only way that biology can be invoked to explain or even justify a hierarchy of human races. However, to the extent that we want to distance ourselves from such views, it can only help to do the work to
depict genetic relationships among humans in ways that allow simultaneously for similarity, diversity, and admixture at the same time as we depict ancestry.

**Postscript exploration, with some open questions**

Figure 13. Regional ancestry after specifying 7 groups. (Source: Li et al. 2009)

The 7 colors in Figure 13 seem to suggest that there are 7 races that correspond reasonably to classical views of races. However, the color coding comes from a genealogical tree on which the colors are superimposed and are given by reference to classical or geographic views of human groups (Figure 14). (That is, you get out what you put in.)
Figure 14. Phylogeny of 51 populations represented in Figure 13. (Source: Li et al. 2009)

If, instead, we divided human genetic variation according to branching distance from the external reference population (namely, chimpanzees) and asked for 7 groups, they could be San, Mbuti Pygmies, Biaka Pygmies, Bantu, Mandenka, Yoruba, and the rest of the human population in the world all put together. (The assumption here is that time
since diverging from a common ancestor is, in molecular clock fashion, reflected in overall genetic divergence.) That would give us a colored belt with 6 narrow bands of color on the left and one solid band stretching out the rest of the way to the right with a few stripelets from the six African groups appearing in that solid band.

Now, someone could sample more in the African groups and produce a tree that, say, splits the San the Mbuti Pygmies, and the Biaka Pygmies and then, to make the 7th group, lumps the rest of the human population in the world all put together. Or someone could sample fewer African groups and end up with, say, three African groups, Mozabite, Bedouin, Palestinians, and the rest of the human population in the world all put together in the 7 human groups.

This sensitivity to sampling leads to the question: Is there a basis for subdivision that is not susceptible to who is sampled more or less and does not depend on color coding that mostly defines the groups before doing the analysis? Perhaps the Principal Component Analysis (PCA) in figure 14 is what’s needed—it uses the full set of information about genetic variation in a sample of 900+ individuals to spread the sampled populations across two dimensions.
That diagram suggests (to my eye) 4 subdivisions of human genetic diversity — African (red), Europe + Middle East + Central/South Asia (all in one group, brown, green, and light blue), E Asia + America + some CS Asia (gold, purple), Oceania (deep blue). Does that bring us back to Africans as one race, albeit with Europeans subsumed in a large mix? That is, does it undermine a reading of the Campbell-Tishkoff diagram that suggests that, if there were 4 races, 3 would be African and one would be African plus the rest of the world put together?

At the same time, keep in mind that, in Li et al’s (2009) analysis, 89% of variation is within populations, 2% is among populations, within groups, and 9% is among groups. (This affirms results dating back to Lewontin 1972, but affirmed by subsequent work, such as Hofer et al. 2008, that, on average, for any genetic locus roughly 5/6 of the variation is within a population, 1/12 is within a region, and only 1/12 occurs among regions.) What such variation means is that it is difficult, on the basis of a random
selection of genetic loci, to assign an individual to one branch or the other. We need to say difficult, not impossible; merely subject to more errors than to correct assignments.

The question I have is why this difficulty is not evident in the PCA plots. Those plots make it look like within groups variation is somewhat less than among groups variation. Perhaps this is because the PCA uses many loci and, contra what I just said above, a combination of a random selection of loci does allow us to discriminate among something like the classical groups? (Remember, however, that people of recent mixed ancestry tend to be eliminated as subjects in these studies.) Perhaps it is because the PCA is biased towards the particular loci that allow us to trace ancestry. Researchers who understand more about what the methods are doing may be able to help here. My current explorations are as follows:

1. Scale the diagram so that variation on the plot is proportional to variation accounted for by each of the first two principal components. (I also rotated it because I like the convention of the major axis of variation being left to right.)

![Figure 15. Scaled and rotated version of Figure 14.](image)

2. Consider, as a thought experiment, groups separated on the first axis. That is, 3 subdivisions of human genetic diversity — African (red); Europe + Middle East + most Central/South Asia + Oceania (all in one group, brown, green, light blue, and deep
blue); and E Asia + America + some CS Asia (gold, purple). Then choose some place on the second PC as if the variation in that direction were all the variation not accounted for by the grouping (rather than actually only 3/5 of it). There's a lot of overlap among the three groups for any position with PC2 > 0.2. This shows how a randomly selected gene (or combination of genes) not captured by the first axis won't be a reliable basis for separating groups; members of two or more groups will share that gene.

3. Now consider groups separated by using both PC axes and imagine choosing a gene (or combination of genes) along the direction of the remaining variation (20% of the total). Again, a randomly selected gene (or combination of genes) not captured by the first two axes won't be a reliable basis for separating groups; members of two or more groups will share that gene.

4. Granted, a randomly selected gene (or combination of genes) captured by the first two axes will do OK in separating groups. However, if the trait we are concerned with involves many genes (not to mention environmental factors interacting over a developmental sequence), we will expect it to be difficult to link differences between any two individuals in different groups to genetic differences.

5. Of course, if the genes that do allow us to separate groups (either in an ancestry method or using PCA) had been the focus of natural selection in divergent environments, then the separation on the ancestry tree or PCA plot would mean something. Is there any evidence for that? Indeed, what would be required to establish evidence for that?
References


