

Transcript: Can we be categorised by our DNA?

Narrator:

This video demonstrates why human genetics doesn't conform to traditional notions of racial groups, and why ancestry testing can't give you the exact results you expect.

By using the analogy of a colour palette.

First of all, geneticists don't use the word 'race' anymore. This is because it carries a huge baggage of misconceptions and misuse.

The term 'ethnicity' or 'ethnic group' is often used, and can mean continental, sub-continental, or regional population groups.

We tend to think of ethnicities as distinct, discrete genetic groups. But humans are rarely stagnant and isolated, we flow, spread, and mix.

This genetic continuity means that the genetic landscape is actually formed of gradients, shaped by geographical features.

The colours red-blue-green are very different and distinct, like we presume different ethnic groups are.

But it is wrong to think that the worldwide genetic landscape is formed of such discrete groups.

A far better analogy is the continuous colour gradients on a colour palette.

The continuity means any categorisation to assign ethnic group labels is subjective and arbitrary. Say I gave you the entire colour palette and asked you to label colours.

How many colour names would you use? Only the primary colours? Secondary colours too? All common colour names?

Not only that; to assign labels to this entire space, you have to cut it into groups by drawing boundaries. Now where exactly do you draw the boundary between red and magenta?

Do you add another category called fuchsia in the middle?

Surveys show this is very subjective, and answers can vary a lot, with different people putting the same colour into different groups, or they use different names.

Here's one survey result example.

This is the same with population groups and labels.

Still, separating major colours is easy. But the more fine-scale you try to get, the more difficult and ambiguous it gets.

It is easy to identify and distinguish between red-green-blue, but can you identify and distinguish between the colours auburn, carnelian, and Persian red?

Back to genetics, take the three connected continents Europe, Asia, Africa.

The 1000 Genomes project started by sampling some population groups from West Europe, Central Africa and East Asia.

The 3 sets of continental populations are easy to separate in the genetic data. These groups, being far apart and thus rather distinct, are different in appearance as well.

That's how early explorers or anthropologists, upon seeing people from far-away lands, thought them to be very different to themselves, and called them distinct races.

But for a genetic analysis, it is very important to have good geographic coverage, which means dense sampling in the reference set.

Consider the huge gaps in 1000 Genomes Phase 1, such as Middle East or South Asia. If using this data as a reference, you'd only have those 3 groups.

If that set is used as reference to analyse Middle Eastern people, they would be marked as mixed with contributions from all 3 continents, when in reality they're an ethnic group of their own, with unique genetic variations.

So, just like the colour palette, it is far easier to assign labels to the extreme ends of a gradient.

For example, in the Europe-Asia horizontal gradient, people of West Europe are often considered the most 'typical' European. This is how we think informally.

But assigning fine-scale labels, such as in the middle of a gradient, is very difficult and ambiguous. Our tendency is to use political boundaries.

But these change, and often split one population into multiple groups, which increases

the ambiguity and numerical uncertainty.

For example, a Northern Irish person could be genetically closer to an average person from Ireland than to an average UK person, but at a country-level grouping they would have to be included in the UK reference set.

So for precision, if you do decide to split the UK into major administrative regions, where do you stop?

England-Wales-Scotland-Northern Ireland? What about the Orkney residents who are genetically quite distinct from Scotland?

Do you go down to county level? Making 109 groups?

Giving hundreds of small groups in an ancestry report would be too much for a user, and impossible to estimate accurately.

While only giving continental groups wouldn't be accurate enough. The groups ancestry testing companies use are a compromise, and have limited precision.

As an example, when analysing Iberian people, even with the latest algorithms and a dense sampling, we found it hard to separate North-West Spain (Galicia) from Portugal, and had to group them in a single reference cluster, to do reliable ancestry estimation in other people.

So, too dense coverage means a problem of plenty. How can we reasonably combine the population groups into something useable?

To have an interpretable number of ethnicity components, and decent precision in estimating the contribution from each? The answer is, we can't!

Remember, this analogy of genetic gradients with a colour palette isn't perfect. Some major aspects of genetics are not reflected.

Populations change over time; migrations and intermixing are two major factors. Let's not forget populations consist of people who are diverse themselves.

[Text on screen]

Narrated by Sofia Maruzza

Produced by Sas Amoah

Research by Dr. Kaustubh Adhikari

Animated by Gwri Pennar