Ancient DNA Suggests Steppe Migrations Spread Indo-European Languages[1]

DAVID REICH

Professor of Genetics, Harvard Medical School
Investigator, Howard Hughes Medical Institute
Senior Associate Member, Broad Institute of MIT and Harvard

Thank you for the opportunity to speak to this group. It is a particular pleasure to speak after Andrew Garrett. I'm going to talk about my work in ancient DNA, which is an extraordinary new enterprise that has become possible on a genome-wide scale since 2010 and has made it possible to take a direct look at past archaeological cultures. We can take skeletons like the one in Figure 1, extract DNA from them, effectively sequence their genome, and compare them to individuals from other ancient archaeological cultures and people today to determine how they are related.

For the first time, we can trace movements of people and ask whether transformations in material culture in the past correspond to movements of people or communication of ideas. It's a revolution in our ability to understand the past.

Figure 2 shows a tree reconstruction of the relationship of Indo-European languages. It is different in some of its features from some of the trees that Andrew showed, and I am not making any claim that this tree or another is correct in all its details.

The relationship among Indo-European languages was articulated at the end of the 18th century by Sir William Jones working in colonial India who noticed, as Andrew said, the connections between Sanskrit and the ancient European languages. Language provides compelling evidence of a strong cultural connection among these disparate places. It's a key observation because it indicates that there was cultural continuity across these regions.

Figure 3 shows how there has been an explosion in the amount of ancient DNA since 2010. There was a trickle of genomes between 2010 and 2013. Suddenly in 2014 and then again in 2015 there was a rapid increase in the number of individuals with genome-wide data. At the end of 2015 it was more than 300; it's now over 1,000. The sample size

1 Read 29 April 2017 as part of the *Indo-Europeanization of Europe* symposium.

FIGURE 1. Corded Ware skeleton. Courtesy of Wolfgang Haak.

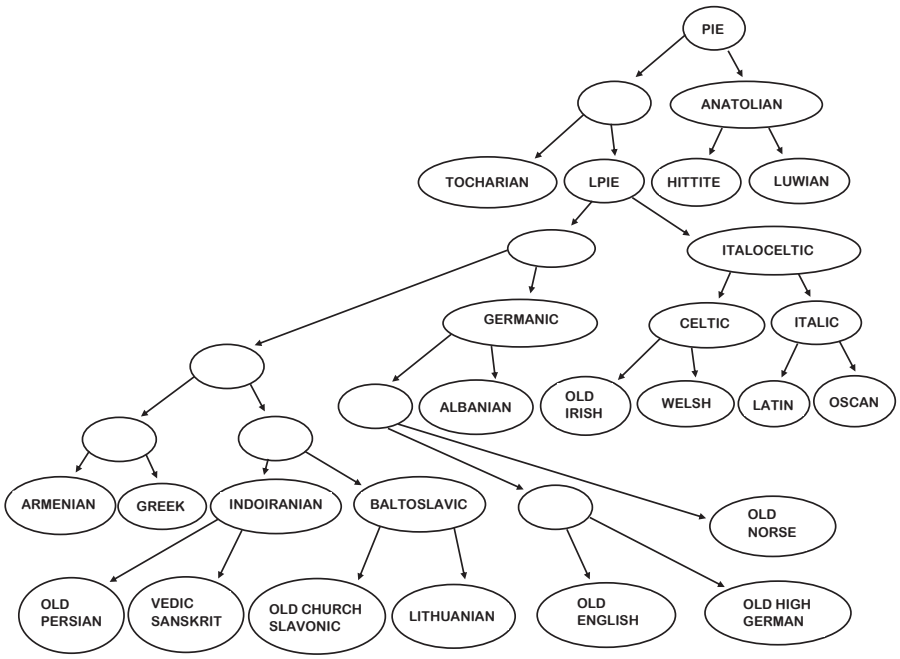

FIGURE 2. Linguistic tree for Indo-European languages. Data from Don Ringe, Tandy Warnow, and Ann Taylor, "Indo-European and Computational Cladistics," *Transactions of the Philological Society* 100, no. 1 (2002): 59–129.

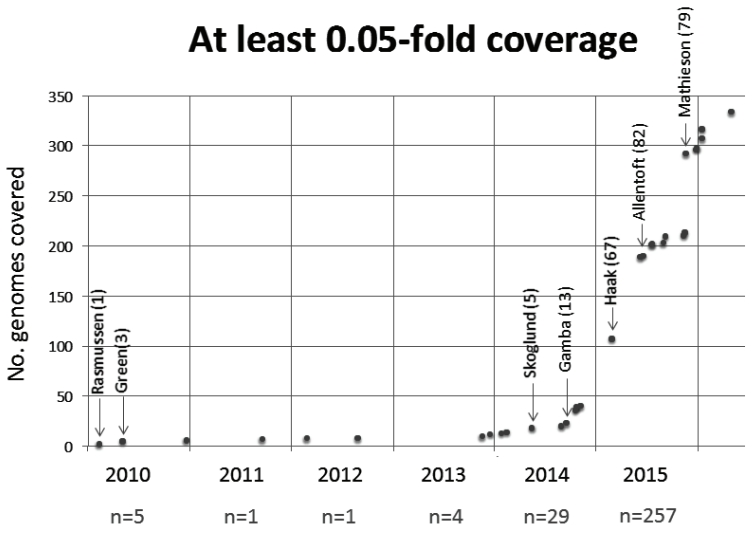

FIGURE 3. Moore's Law of Ancient DNA.

is rapidly increasing and this is allowing us to study how people from the past relate to each other and those of the present.

In 2009 a very important study was published in which mitochondrial DNA—in the energy factories of cells—was successfully extracted and sequenced from ancient Europeans. Bramanti and colleagues compared the mitochondrial DNA of hunter-gatherers and farmers in Europe who lived before and after about 8,000 years ago when farming arrived in the region. They noticed that the mitochondrial sequences were almost entirely different between hunter-gatherers and farmers, suggesting that new people had arrived along with the new economy.

In 2012 whole genome data arrived, covering about 200,000 times more DNA letters. That study compared hunter-gatherers and farmers who lived in southern Sweden 5,000 years ago. The farmers were not at all like present-day people from Sweden, and instead were much more similar to present-day people from Sardinia, an island off southern Europe. The hunter-gatherers were not like present-day people from Sweden either. The authors proposed a model in which a mixture of these two populations, in different proportions, formed Europe today (Figure 4).

But this model isn't a full description of the main patterns of variation in Europe. In that same year (2012), Nick Patterson, a colleague who I've worked with for 16 years, found something that didn't seem to be consistent with these observations. To understand what Patterson found, I need to provide a bit of background about the genome and

FIGURE 4. Estimated Neolithic farmer ancestry (red) in European populations. Ancient DNA shows the advent of farming was accompanied by large-scale migration from the Near East. From Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M. Thomas P. Gilbert, Anders Götherström, and Mattias Jakobsson, "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe," *Science* 336, no. 6080 (2012): 466–69. Reprinted with permission from AAAS.

how we can study it to learn about the deep human past. The genome is a sequence of about 3 billion paired chemical units that can be thought of as letters—adenine (A), cytosine (C), thymine (T), and guanine (G)—that are almost always the same between any two genomes, but occasionally are different. Between any two copies of a human genome, there is typically about one difference in every thousand positions. That's about 3 million differences between these two copies of the genome, enough to learn a lot about how those differences occurred over time.

Patterson looked at about 600,000 positions in the genome that are commonly observed to be different among people, and a set of about 50 worldwide populations. For each of the about 50 populations, he tested all other possible pairs of populations to see if there are population pairs for which the test population looks intermediate, as would be expected for a mixture.

Patterson found a strong, highly significant signal showing that the French (and other populations of northern European ancestry) tended to be genetically intermediate between two other populations. One was always Sardinians, who we now think of as descendants of the first farmers of Europe. But the other population was, amazingly, Native Americans. It was definitely Native Americans who were giving the

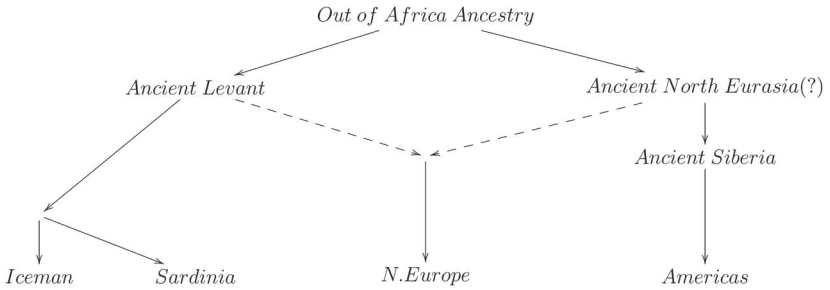

FIGURE 5. "The Ghost of North Eurasia," the third ancestral population for present-day Europeans. A proposed model of population relationships that can explain some features observed in our genetic data. Reproduced by permission from Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich, "Ancient Admixture in Human History," *Genetics* 192, no. 3 (2012): 1065–93. Courtesy of the Genetics Society of America.

strongest signal—the signal was stronger when Native Americans were used in the analysis than when East Asians or Siberians were used. The only way this could occur was if Northern Europeans were a mixture of populations related, perhaps distantly, to Sardinians on the one hand and Native Americans on the other.

To explain these perplexing observations, we proposed a new population called the "Ancient North Eurasians." This was a "ghost population" from which we didn't have any samples, and which doesn't exist anymore, but must have existed more than 15,000 years ago (Figure 5). Descendants of this group contributed both to Native Americans and to northern Europeans.

A year later (Figure 6), this ghost population was found. Eske Willerslev and colleagues working in Denmark obtained DNA from a boy who lived about 24,000 years ago around Lake Baikal in Siberia. It perfectly matched this predicted source population. It's in fact a better match than the Native Americans. This is a recurrent theme in ancient DNA research—predict ghost populations statistically from samples we have and then find them. In some cases we haven't yet found them, but we will. It's a very exciting thing to be able to do.

How does the finding of a mixture in northern Europeans of people descended from Ancient North Eurasians and Sardinians relate to the evidence of mixtures of European hunter-gatherers and European farmers? The answer is that both these mixtures occurred. Europeans today are the result of a mixture of not two but three very different ancestral populations.

In 2014, my colleagues and I obtained high-quality DNA from a hunter-gatherer from 8,000 years ago and a farmer from 7,000 years

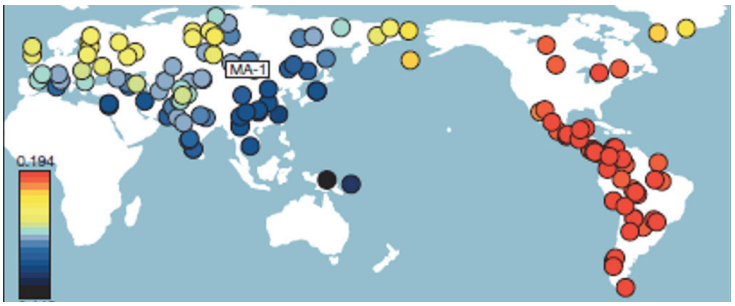

FIGURE 6. "The ghost is found." Third ancestral population for present-day Europeans. Sample locations and MA-1 genetic affinities heat map. Reprinted by permission from Springer Nature: Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen et al., "Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans," *Nature* 505, no. 7481 (2014): 87–91. Copyright © 2014.

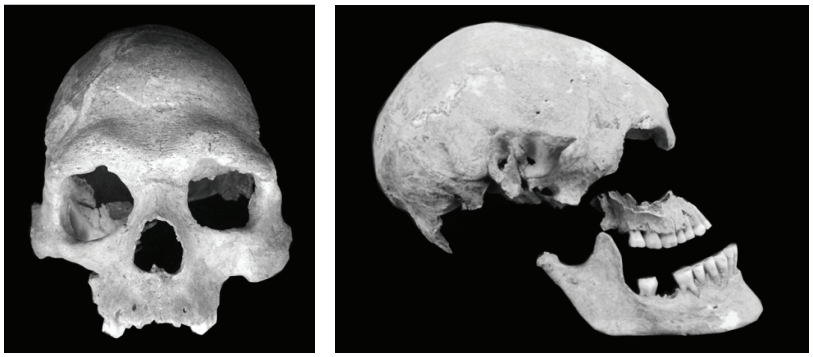

FIGURE 7. High-quality genomes from Ancient Europeans. Left: Luxembourg, Loschbour hunter-gatherer ~8kya; Right: Germany, Stuttgart farmer ~7 kya. Reprinted by permission from Springer Nature: Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant et al., "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans," *Nature* 513, no. 7518 (2014): 409–13. Copyright © 2014.

ago (Figure 7). To put these ancient samples into the context of present-day people, we assembled data from 777 present-day West Eurasians from whom we had data at approximately 600,000 positions in the genome. This provided a rectangular matrix that we multiplied by itself to form a 777-by-777 square matrix that revealed to us how closely related each sample was to each other sample. We used the technique of principal component analysis to summarize the differences across the samples.

What we found was a dramatic pattern whereby West Eurasians formed two parallel lines (Figure 8). The left one contained essentially

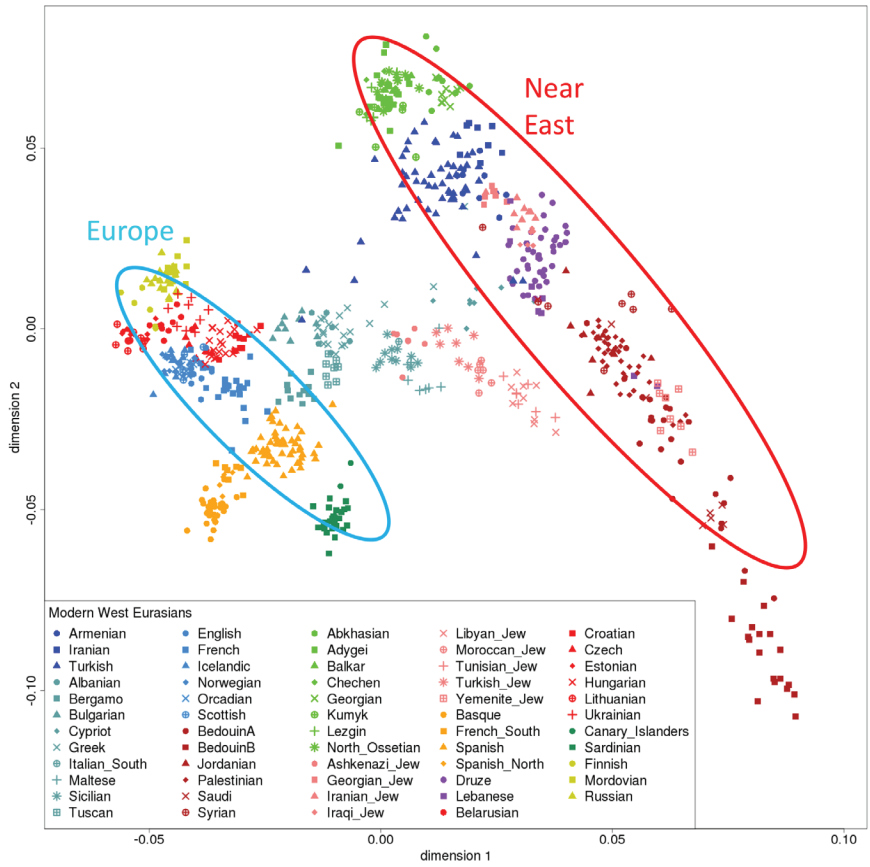

FIGURE 8. Principal component analysis.

everyone in Europe. The right one contained essentially everyone in the Near East (ranging from the Levant up to the Caucasus). The populations in between were in areas of plausible contact between Europe and the Near East, mostly island Mediterranean groups as well as Jewish groups.

This qualitative observation suggested to us that two things might be going on. The big gap between the European and Near Eastern lines might correspond to mixture into the ancestors of present-day Europeans of ancient European hunter-gatherers—a group that did not contribute to Near Easterners. But there's also the perpendicular dimension, and perhaps that's related to the pattern we see in Native Americans.

Figure 9 shows a model of history consistent with the data. Walking through the figure, the Mbuti are hunter-gatherers from central Africa

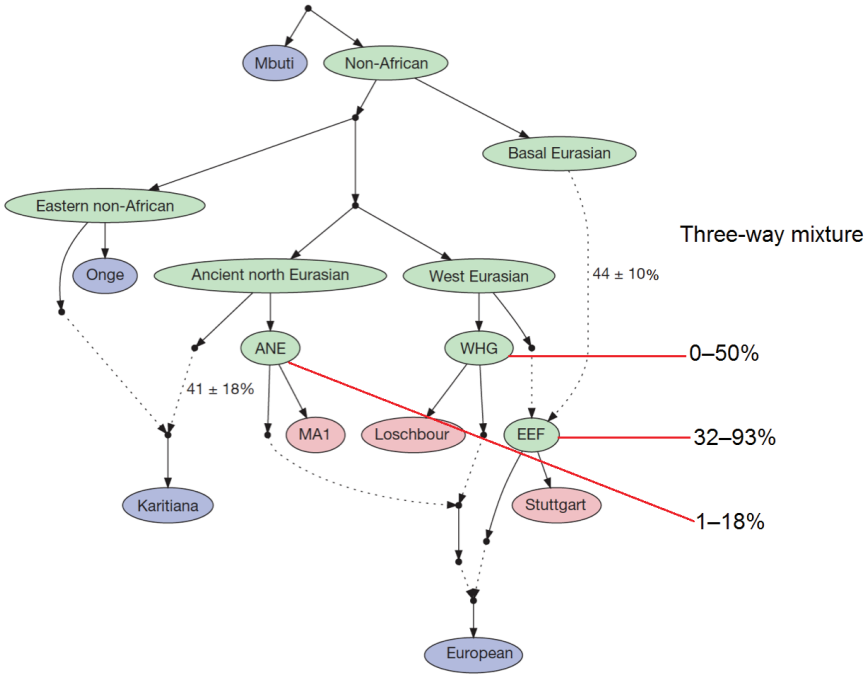

FIGURE 9. Model of history consistent with data. Reprinted by permission from Springer Nature: Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant et al., "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans," *Nature* 513, no. 7518 (2014): 409–13. Copyright © 2014.

that we use as a reference point that is symmetrically related to all non-African populations. The Onge are isolated hunter-gatherers from Southeast Asia. "MA1" is an "Ancient North Eurasian": the 24,000-year-old boy from Lake Baikal. "Loschbour" is the 8,000-year-old European hunter-gatherer. "Stuttgart" is the 7,000-year-old European farmer.

The Onge share mutations at an equal rate with Ancient North Europeans and with Loschbour, showing that these two groups are consistent with descending from a common ancestral population that split earlier from the ancestors of East Asians.

Native Americans are a mixture of Ancient North Eurasians and an ancient East Asian group, which explains why Native Americans have an affinity to East Asians but also an affinity to Europeans. Native Americans are a mixture of about one-third ancestry related to this ancient North Eurasian group and another two-thirds related to East Asians.

The first European farmers represented by "Stuttgart" are a mixture of a group related to European hunter-gatherers and some deeply splitting lineage which is another ghost population that we still haven't found but which we predict existed. I am confident it will be found.

Europeans today are a mixture of three sources: Ancient North Eurasians, European hunter-gatherers, and European farmers, in proportions that vary across Europe. The largest proportion of ancestry is from the first farmers but large proportions also derive from these other groups.

How did this new ancestry get there? The data that we had in 2014 didn't have any of the Ancient North Eurasian ancestry. It was all consistent with a simple mixture of hunter-gatherers and farmers. But today this ancestry is everywhere. How did this transformation happen?

In 2013, a study was published—again of mitochondrial DNA—that gave a clue (Figure 10). This study looked at more than 300 samples from nine successive European archaeological cultures, beginning from hunter-gatherers all the way to people who lived around 4,000 years ago. My laboratory's contribution was to develop a test for continuity, which assessed whether successive populations were consistent with being just a random sample from previous populations without additional immigrations. Figure 11 shows a discontinuity between hunter-gatherers and farmers, no substantial discontinuity between the first few farming cultures, and then a major discontinuity around 4,500 years ago in association with an archaeological phenomenon known as the Corded Ware culture. This was a suggestion that some new ancestry came into central Europe 4,500 years ago around the time of the Corded Ware culture.

Two years later we published a whole genome study that included a series of samples from the same region of Germany and additional data from Hungary, Spain, and Russia. Altogether we analyzed 94 samples to study how changes occurred (Figure 12).

Figure 13 is a reprise of Figure 8—two parallel lines, Europe and the Near East—showing how the ancient samples plotted over time. The hunter-gatherers fall beyond Europe in the direction of European differentiation from the Near East as expected. At the left (orange diamonds) are Scandinavian hunter-gatherers from Sweden, and below them are western European hunter-gatherers from Spain and from Luxembourg. Above both of these are eastern European hunter-gathers from Russia. So this gradient is already present amongst the hunter-gatherers.

The first farmers from Anatolia—but also from northern and western Europe—fall in a very different place on the plot, close to

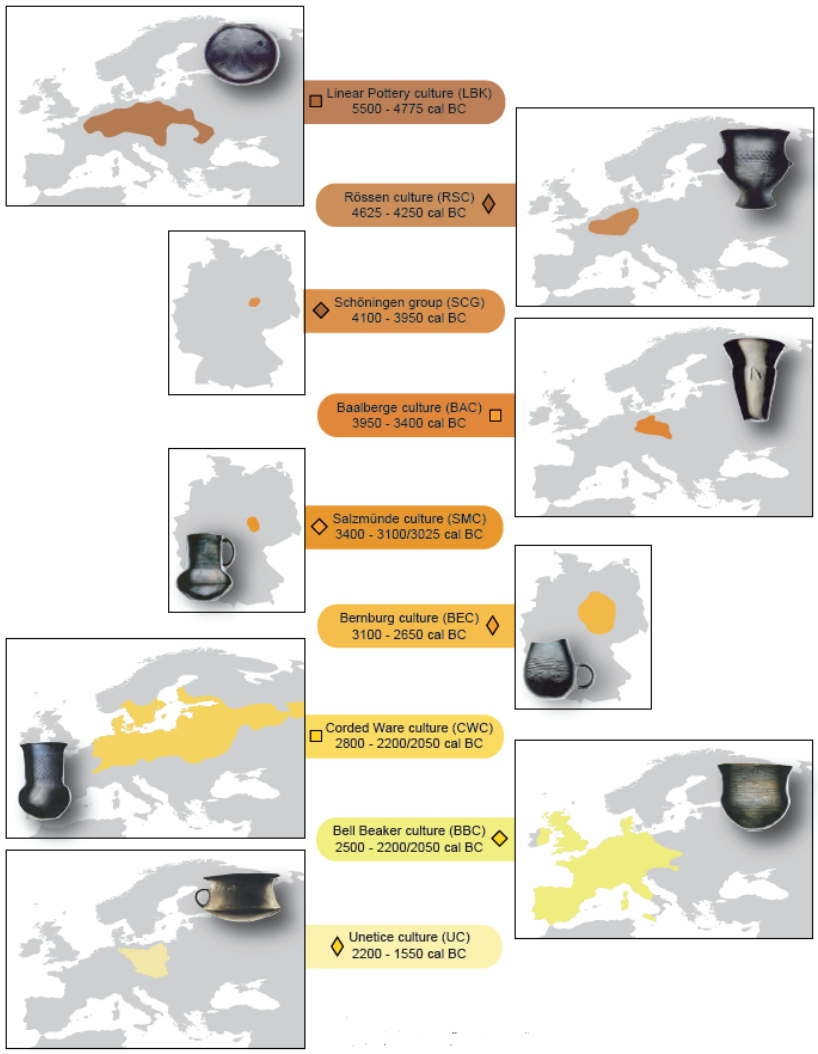

FIGURE 10. Evidence from mitochondrial DNA. Photos of ceramic vessels: © State Office for Heritage Management and Archaeology Saxony-Anhalt; Juraj Lipták. From Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker et al., "Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity," *Science* 342, no. 6155 (2013): 257–61. Reprinted with permission from AAAS.

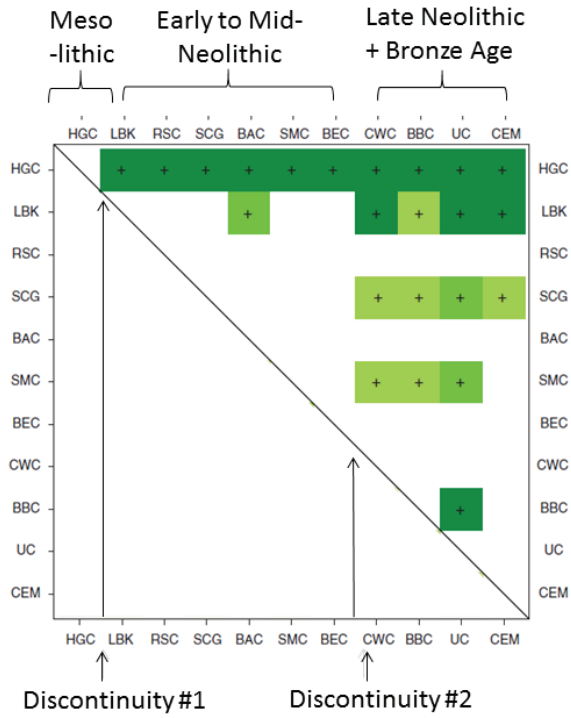

FIGURE 11. When did the second population turnover occur? From Guido Brandt, Wolfgang Haak, Christina J. Adler, Christina Roth, Anna Szécsényi-Nagy, Sarah Karimnia, Sabine Möller-Rieker et al., "Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity," *Science* 342, no. 6155 (2013): 257–61. Reprinted with permission from AAAS.

present-day Sardinians. It's quite clear that a population related to Anatolian farmers arrived 8,500 years ago and was responsible for bringing this new ancestry to Europe. Sardinians fall in a very similar place, consistent with the idea that Sardinians are isolated descendants of these first farming populations.

The next thing that happens among European farmers is that they show a leftward shift in the plot (Figure 13), reflecting mixture with local hunter-gatherers who have not yet been absorbed. We see this again and again in ancient DNA: the arrival of a migrating population, which then takes hundreds or even thousands of years to admix with local groups.

Meanwhile, in Eastern Europe, a population forms that archaeologists call the Yamnaya. The Yamnaya from whom we have data are from Far Eastern Europe close to the Ural Mountains. They're halfway

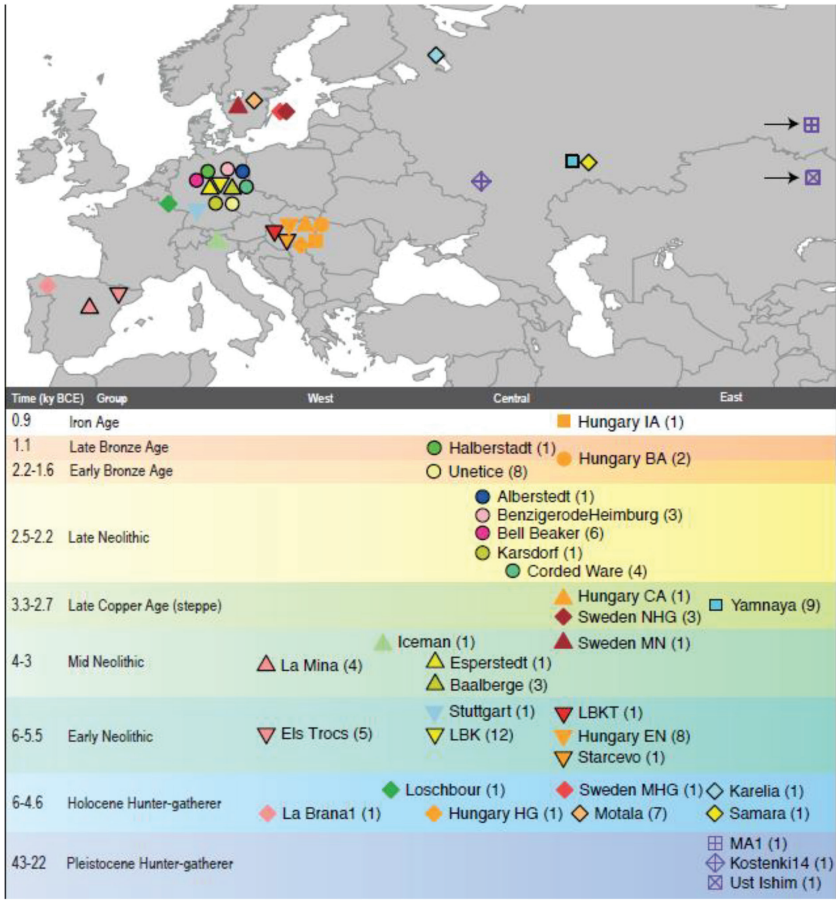

FIGURE 12. Genome-wide data from 94 ancient Europeans. Reprinted by permission from Springer Nature: Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt et al., "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe," *Nature* 522, no. 7555 (2015): 207–11. Copyright © 2015.

between the northern tip of the Near East and the hunter-gatherers of eastern Europe on the plot, and that reflects real mixture events. But populations like those we see in Europe today have not yet formed. Then suddenly they form 4,500 years ago, beginning with the people of the Corded Ware culture.

Figure 14 summarizes the evidence in another way. The horizontal orange bars at the top show that the first farmers of Europe had almost all their ancestry from Anatolian farmers. Over time more hunter-gatherer ancestry came into these populations as they mixed with local hunter-gatherers represented in the figure in blue. But beginning around

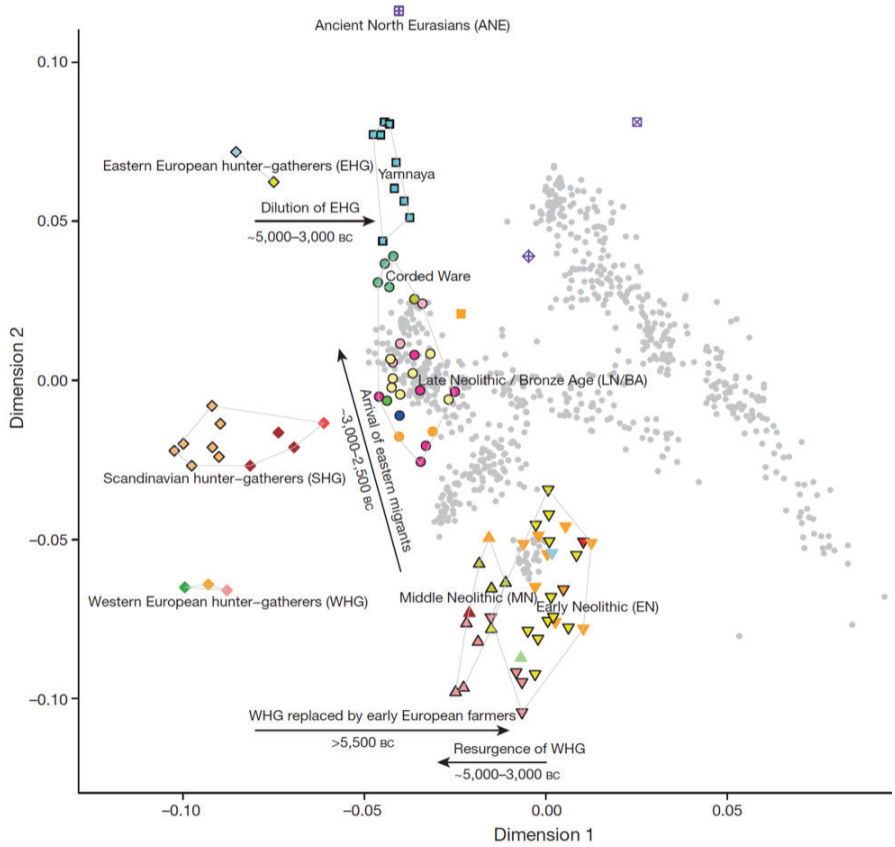

FIGURE 13. Time series of population turnovers. >5500 BCE Hunter Gatherers; 5500–4000 BCE Near East Migration; 4000–3000 BCE Hunter-Farmer Mixture; 4000–3000 BCE Yamnaya Steppe Pastoralists; <3000 BCE Formation of Modern Europe. Reprinted by permission from Springer Nature: Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick Bastien Llamas, Guido Brandt et al., "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe," *Nature* 522, no. 7555 (2015): 207–11. Copyright © 2015.

4,500 years ago, with the Corded Ware culture, a new ancestry from the steppe arrived (shown in green). It was a massive replacement: at least 70 percent of ancestry in Corded Ware culture people. In the German population and forever afterward in northern Europeans there was a large proportion of ancestry from this group. This is the single most important contributor to northern European populations today.

A summary (Figure 15) is that the population of Europe was in the last 10,000 years massively transformed by two major migrations. After 8,500 years ago a mass movement of farmers mixed with local

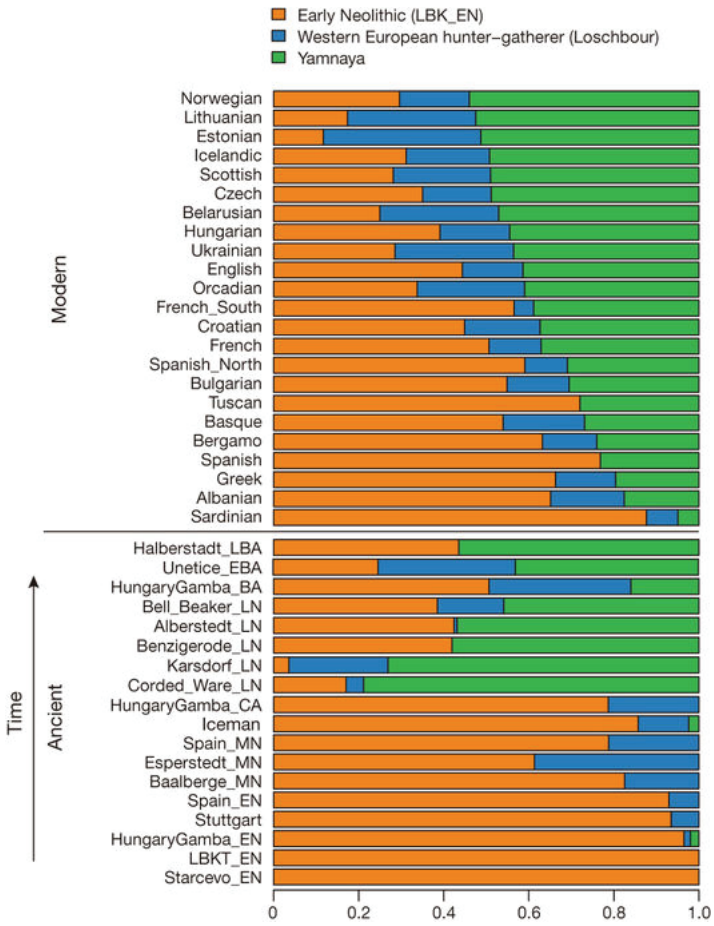

FIGURE 14. Population changes after advent of agriculture. Steppe-farmer mixture <3000 BCE. Reprinted by permission from Springer Nature: Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt et al., "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe," *Nature* 522, no. 7555 (2015): 207–11. Copyright © 2015.

hunter-gatherers, and after 4,500 years ago a mass movement of steppe pastoralists mixed with the established mixture of hunter-gatherers and farmers. Today Europeans are largely a mixture of these three ancestral components.

I now want to back up and connect this to the Indo-European language puzzle. Ancient DNA has falsified the single-strongest argument for the theory that farming spread from Anatolia, first suggested by Colin Renfrew. Renfrew thought massive movements of people were

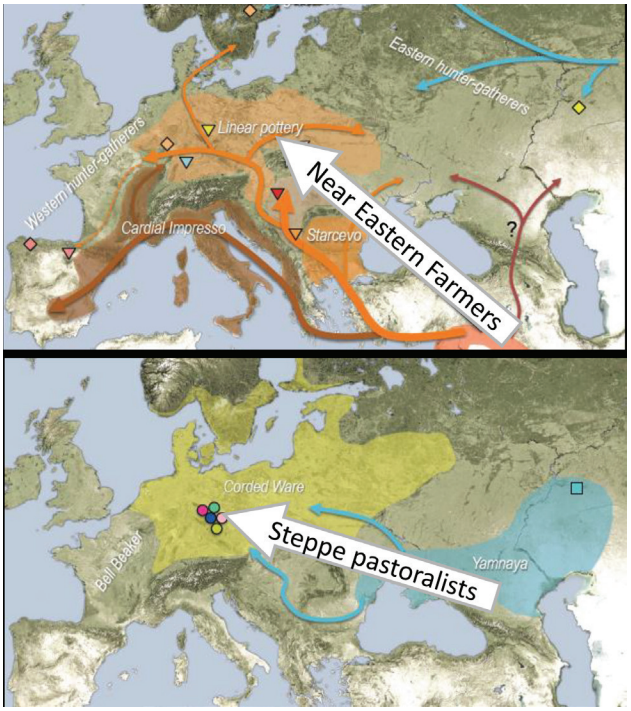

FIGURE 15. Population transformation by two migrations. Migration 1: 6500 BCE, Near Eastern farmers; Migration 2: ~2500 BCE, Steppe pastoralists; Implication: Some Indo-European languages likely came with second migration. Reprinted by permission from Springer Nature: Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt et al., "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe," *Nature* 522, no. 7555 (2015): 207–11. Copyright © 2015.

rare in history, and that the only plausible one into Europe in the last 10,000 years would have been the spread of farmers from Anatolia, because they had an economic lifestyle advantage compared to the hunter-gatherers. But once the densely settled farming population was established in the region, it would have been hard for anybody else to make an impact. Similarly in India, the British and Mughals were in political control for hundreds of years but hardly made a demographic impact. But genetics shows that there *was* a massive later movement.

The first culture with the steppe ancestry is the Corded Ware complex. Corded Ware samples from whom we have genetic data are genetically relatively homogeneous and derive most of their ancestry from the Yamnaya. Because they overturned the local population in such large numbers, they almost certainly brought an Indo-European

language ancestral to some spoken in Europe today. So this supports some version of the steppe hypothesis (Figure 16).

I also wanted to speak a little bit more broadly on the question of steppe-related ancestry in India. Steppe ancestry is confidently inferred to exist in almost every Indian group of Indo-European speakers without exception (Figure 16). It is higher in groups that speak Indo-European languages consistent with the idea of a Yamnaya affinity to Indo-European languages. There is particularly interesting evidence from the Y chromosome, which boys inherit from their fathers who inherit it from their fathers. This shows that a large fraction of the male ancestry of India is related to that in Eastern Europe within the last 5,000–7,000 years.

And yet popular models of the steppe hypothesis of Indo-European language origins also do not seem compatible with the data in some ways, as shown in Figure 17. Often in the story told with the steppe hypothesis, the Yamnaya give rise to later groups called the Sintashta and Andronovo that then contributed to India. But genetic data from those populations so far call into question that model, as they do not work statistically as sources of ancestry in India. But perhaps with more sampling in central Asia we will find better matches to the source population.[2]

So there is now compelling evidence that the spread of the Yamnaya archaeological culture was the vector that also spread all the Indo-European languages spoken today. Genetics has made some important progress toward solving the more than two hundred year old Indo-European language puzzle. It will be exciting to work with archaeologists and linguists over the coming years to work out the implications of these new genetic findings.

2 In the year since this lecture was delivered, my laboratory and another laboratory have generated new ancient DNA data from Central Asia. With these data in hand, we have been able to show that groups like Sintashta and Andronovo in fact have mixed with some of these newly sampled populations to produce a plausible source population for South Asians.

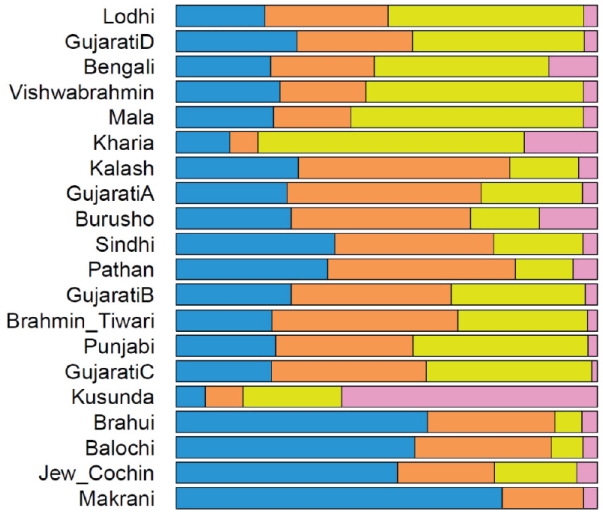

FIGURE 16. Steppe-Related Ancestry in India is present in almost every Indo-European and Dravidian-speaking population, with higher relative proportions in Indo-European speakers. Orange = Steppe-related; Yellow = Iranian-related. Reprinted by permission from Springer Nature: Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes et al., "Genomic Insights into the Origin of Farming in the Ancient Near East," *Nature* 536 (2016): 419–24. Copyright © 2016.

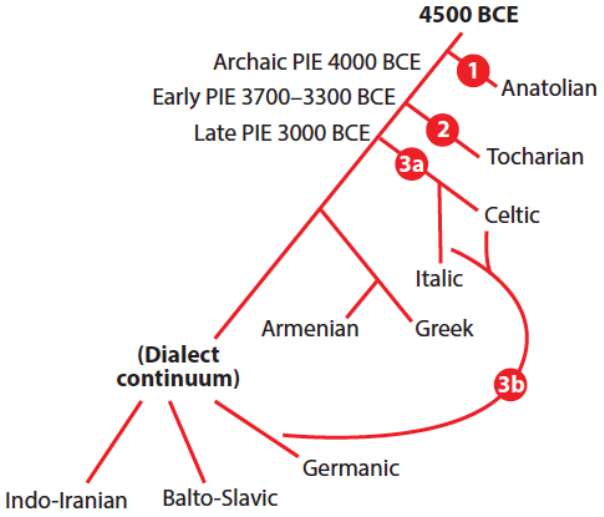

FIGURE 17. Indo-European reconstruction. Yamnaya is a plausible culture responsible for spreading "Late Proto-Indo-European" languages. The ultimate homeland of the group that also spread Anatolian languages is less clear. Reproduced with permission of Annual Reviews via Copyright Clearance Center, from David W. Anthony and Don Ringe, "The Indo-European Homeland from Linguistic and Archaeological Perspectives," *Annual Review of Linguistics* 1 (2015): 199–219.