RECONSTRUCTING THE HISTORY OF LIFE
A GENETIC APPROACH

KRISHNAPRIYA TAMMA

There are 7 billion people, more than 100 trillion ants, and more than a billion honey bees in the world. Where did all this diversity come from? How did we evolve - what's the human story? In this article, the author explores some of the ways in which we've tried to answer these questions.

Humans have always been curious about the origins of life on earth. Many theories and explanations have been offered over the course of human history. When Darwin and Wallace first presented the theory of evolution by natural selection, it led to a great deal of controversy. However, today there is indisputable evidence for this theory. Evolution, according to Darwin, proceeded through '**Descent with modification**', which simply denoted the passing of traits from parent to offspring, albeit imperfectly. Offspring, while being largely similar to their parents, may show some differences. Over time, these differences may be selected and result in a divergence strong enough for a new species to be formed. This new species is still related to the older one, since it is derived from it. Just like my cousin and I share a common ancestor in the form of a grandfather, species too share common ancestors with other related species. Going sufficiently far back in time, we can actually trace the ancestor for all life on earth (about 3.8 billion years ago). Therefore, we can reconstruct the history of life on earth by carefully reconstructing relationships between species.

Phylogenetics is the field of biology that focuses on the reconstruction of relationships between different species, and the reconstruction of evolutionary history of lineages. Such phylogenetic trees or phylogenies (refer Fig.1) can be constructed using morphological or genetic data. But, thanks to advances in the field of genetics, genetic data or DNA-based information has given us greater power to reconstruct relationships, even amongst very closely related and morphologically similar species.

An organism's **genome** is its complete set of DNA, including all of its genes. A copy of the entire genome of an organism – more than 3 billion DNA base pairs in humans – is found within each cell of its body. Each copy of the genome contains all the information needed to build and maintain the entire organism.

DNA, the blueprint of life, is double stranded. Deoxyribonucleic acid is a molecule that carries genetic instructions for growth, development, functioning and reproduction of all known living organisms and many viruses. Most DNA molecules consist of two antiparallel biopolymer strands coiled around each other to form a double helix. Much like proteins are polymers of amino acids, each of the two DNA strands is a polynucleotide, or a polymer of nucleotides. Each nucleotide is composed of a phosphate group, a sugar called deoxyribose, and one of four nitrogen-containing nucleobases – cytosine (C), guanine (G), adenine (A), or thymine (T). The sequence in which these appear within a DNA molecule dictates the sequence of amino acids in proteins. The nucleobases on one strand pair up with bases at corresponding positions on the other strand to form **base pairs**.

A **base pair** in DNA is of two types: A-T or C-G. The nucleotides in a **base pair** are

Figure 1. The structure of DNA.
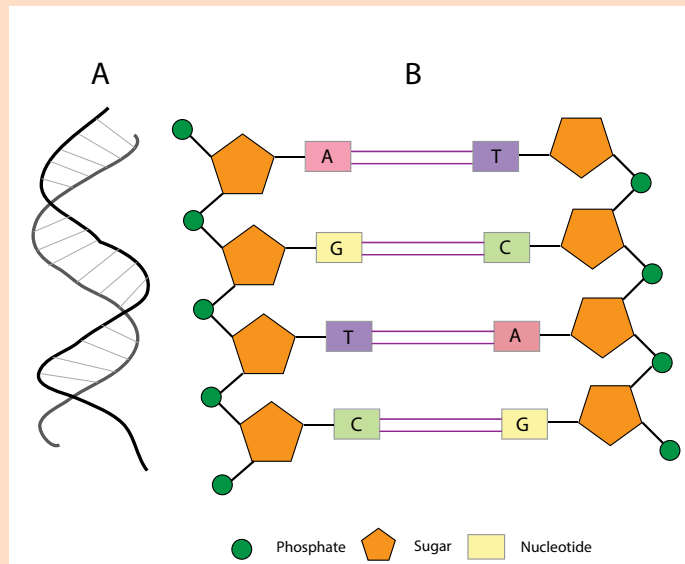
complementary which means their shape allows them to bond together with hydrogen bonds.

A **gene** is a locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity. The transmission of **genes** to an organism's offspring is the basis of the inheritance of phenotypic traits.
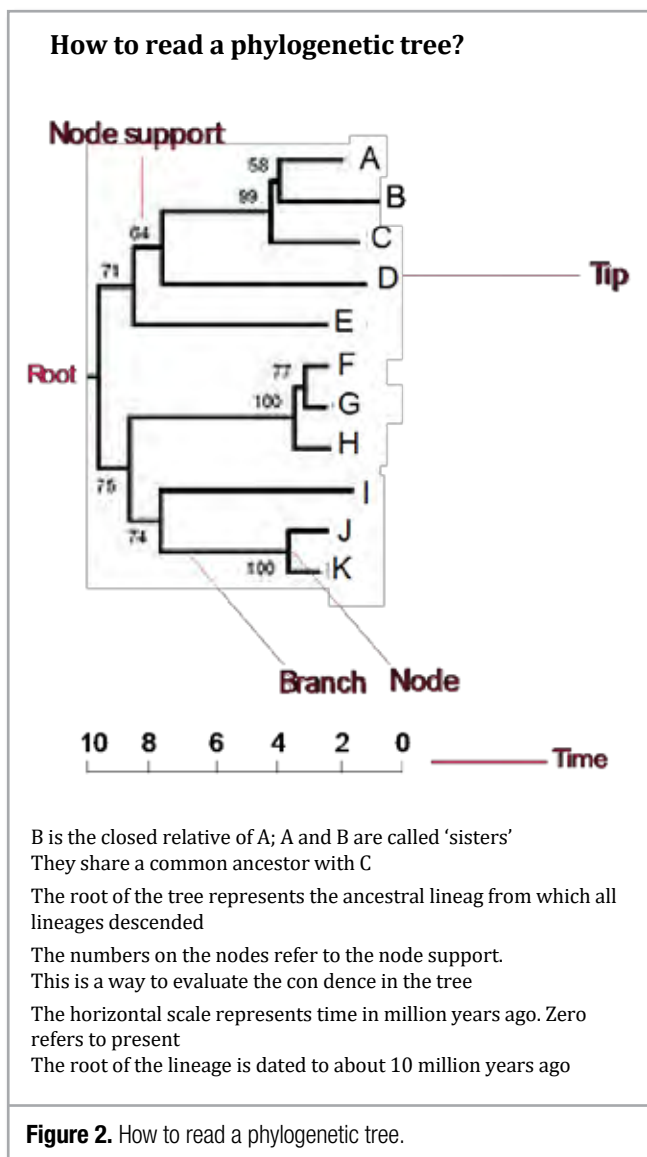
How is this done? By comparing the sequence of nucleotides in the genomes of any two organisms, we can calculate the genetic variation within, and between, the species they belong to. This variation is measured in terms of their genetic distance, and represents the number of bases (A, T, G and C) that are similar (and dissimilar) between the two organisms/species. Through special computer programs, these genetic distances can be used to build a phylogenetic tree of relationships. In general, the genetic distance between two taxa is inversely related to their degree of relatedness. Thus, the value of genetic distance is lower between closely related taxa, and higher between distantly related taxa. This information can be used not just to build relationships between different species, but also to understand how different populations of species are related to each other. Can this tell us anything about the evolution of humans?

How to read a phylogenetic tree

A phylogenetic tree is a diagrammatic representation of the relationships between taxa (Fig. 2). Species that are closely related are closer to each other on a tree, and are linked by fewer branches. Let's break down a phylogenetic tree to understand this better. The first thing that we notice is the overall shape of the phylogenetic tree - the overall branching pattern of the tree is called its 'topology'. At the base of the tree is the 'root'. An out-group, one that is not within the lineage being examined, is necessary to root the tree. The out-group provides a reference point which can help clarify the relationships between species in the 'ingroup' (the one we are interested in) better. The 'tips' refer to the species being compared, and are connected to other tips by 'branches'. The 'internal nodes' refer to the ancestors from whom two or more descendant species

diverge. The nodes represent the point at which the two tips shared a recent common ancestor. From the ancestor, two new species (or generally lineages) are born. Thus, nodes also represent inferred speciation events. Taxa (tips) that share a common ancestor are said to be 'sister' to each other.

For instance, to a casual observer the bonobo and chimpanzee, our primate cousins, look very similar. However, genetically they are quite distinct. In the ape tree (Fig. 3), humans, chimpanzees, bonobos, gorillas and orangutans represent the tips. All apes share a common ancestor that diverged from other primates. While chimps and bonobos are sisters to each other, humans too share a common ancestor with both of them. So that's how we are related to primates! How long ago did this common ancestor between humans and primates exist?

How to read a phylogenetic tree?

B is the closed relative of A; A and B are called 'sisters'
They share a common ancestor with C

The root of the tree represents the ancestral lineag from which all lineages descended

The numbers on the nodes refer to the node support.
This is a way to evaluate the con dence in the tree

The horizontal scale represents time in million years ago. Zero refers to present
The root of the lineage is dated to about 10 million years ago

Figure 2. How to read a phylogenetic tree.

Finding age on the 'tree of life'

Not only does a phylogenetic tree allow us to reconstruct relationships between different taxa, but it also allows us to infer the timing of certain evolutionary events, through a process called 'divergence dating'. Divergence dating is a method based on the molecular clock hypothesis, which suggests that mutations are accumulated over time at some uniform rate. By calculating the number of base differences between the genomes of two species, we can use our knowledge of the rate at which these differences have accumulated to estimate the time when the two lineages split. Although there is no fixed rate at which sequences evolve, we can model the rate as a distribution (say a normal distribution) to estimate divergence (fossils are often used to calibrate the tree since they provide an independent reference age for a clade; ancestor + descendants = clade).

One example of divergence dating can be seen in the primate tree (Fig. 3) which shows us that the chimpanzee and human lineages split about 4-5 million years ago. This means that the human lineage, of which *Homo sapiens* are the sole surviving species, must most likely have evolved around 5 million years ago. Fossil records confirm this. They also show us that although many species of *Homo* evolved between 5 million years ago and now, including *Homo neanderthalensis*; all except *Homo sapiens* have become extinct.

Lineage: A lineage is defined as any continuous line of descent; any series of organisms connected by reproduction by parent of offspring.

Speciation: Speciation describes evolutionary process by which populations become reproductively isolated, eventually leading to the formation of two separate species. Species often diverge from a common ancestor. As they diverge they accumulate mutations, ultimately becoming reproductively isolated – meaning that the individuals of the two groups cannot breed successfully. Reproductive isolation is considered a necessary condition for speciation.

Mutation: Every time a cell replicates, its entire genome is copied. In some cases, there are errors in copying, which result in incorrect nucleotide being inserted into the copies. These changes in the genome are called mutations. Often mutations can change the amino acid that they code for, resulting in changes to proteins.

What if we were able to expand the scope of what we've done with the primate map, and construct a dated tree that shows us how all known species on Earth are related to each other? A project called the Tree of Life aims to do exactly this with molecular data, and using some of the most advanced genetic and computation methods. What is interesting is that Darwin and subsequent evolutionary biologists have largely viewed relationships in the form of a 'tree of life'. But recent advances in our understanding of horizontal gene transfers, involving exchange of genetic material between long-diverged species, make it increasingly likely that all life may actually be related to each other through Web of Life scenarios. This is especially true of prokaryotes.

Do we have all the data we need for completing the Tree of Life? Although we have data today for many more species than ever before, we are still discovering many new species that we did not know of before. Even amongst the species that we know of, there are many from the tropics for which we have no genetic data.
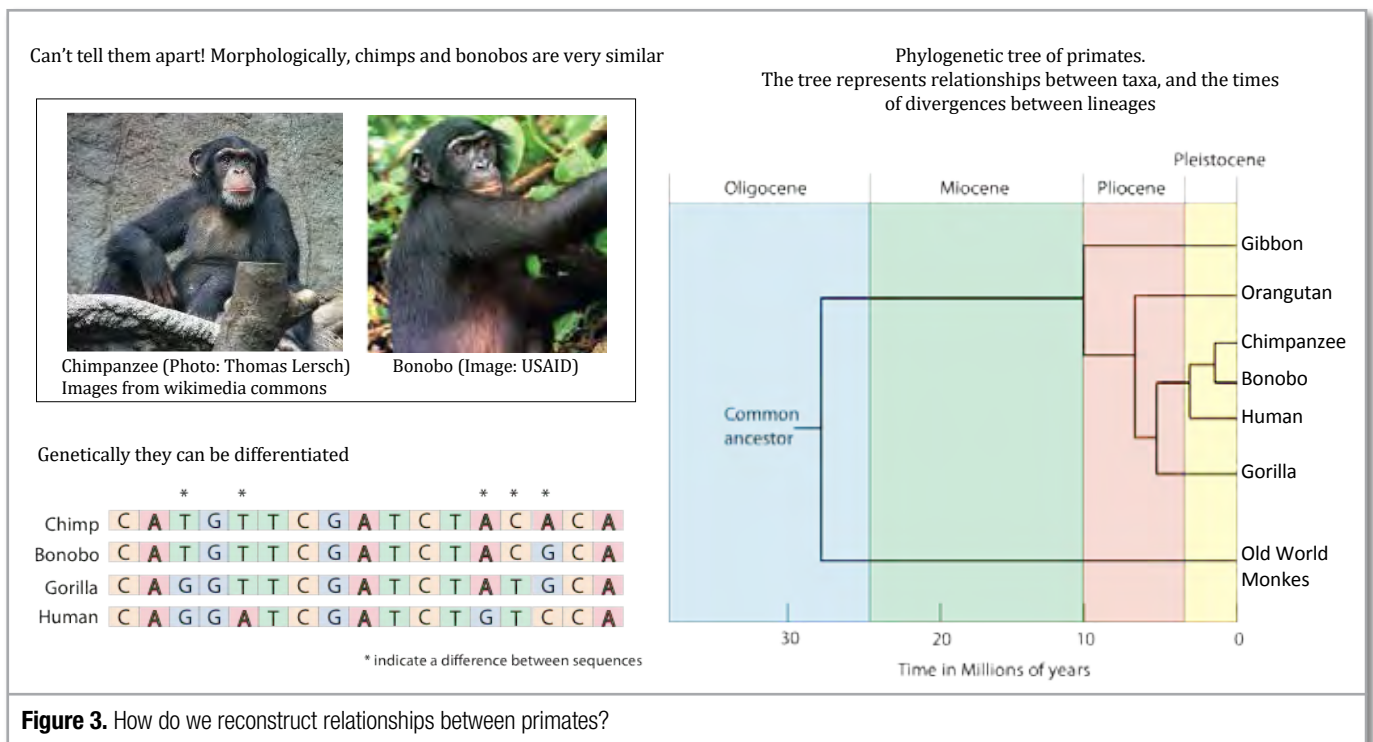
What can we learn from the Tree of life? One can observe startling differences in the distribution of lineages across the Tree of Life, imperfect though it may be. Some clades have a high diversity of lineages, while some others may be represented by a single

Clades: Derived from the Greek word "klados", meaning branch or twig. A clade is a monophyletic taxon; or a group of organisms which includes the most recent common ancestor of all of its members, and all the descendants of that most recent common ancestor.

Prokaryotes: Single-celled organisms that lack nucleus and other membrane-bound organelles.

species. These differences in richness may have been caused both by intrinsic (biological) and extrinsic (historic effects). For instance, small mammal lineages including Rodentia (rats and mice) and Chiroptera (bats) are amongst the richest of mammal lineages. This is probably a result of their small body size and high reproductive rates. Added to this, certain historical factors could have also resulted in rapid evolution and speciation in some groups. For instance, it is believed that the spread of the grasslands in Oligocene (35 million years ago) led to the evolution of hypsodonty (high-crowned teeth) and increase in diversity of animals that showed hypsodonty.

Thus, phylogenies not only help us reconstruct relationships, they also allow us to investigate the mechanisms through which biodiversity is accumulated. In Fig. 4, we can see that some clades are more diverse (have more species) than others,

Can't tell them apart! Morphologically, chimps and bonobos are very similar

Chimpanzee (Photo: Thomas Lersch)
Images from wikimedia commons

Bonobo (Image: USAID)

Genetically they can be differentiated

				*			*								*		*		*	
Chimp	C	A	T	G	T	T	C	G	A	T	C	T	A	C	A	C	A
Bonobo	C	A	T	G	T	T	C	G	A	T	C	T	A	C	G	C	A
Gorilla	C	A	G	G	T	T	C	G	A	T	C	T	A	T	G	C	A
Human	C	A	G	G	A	T	C	G	A	T	C	T	G	T	C	C	A

* indicate a difference between sequences

Phylogenetic tree of primates.
The tree represents relationships between taxa, and the times of divergences between lineages

Oligocene   Miocene   Pliocene   Pleistocene

Gibbon
Orangutan
Chimpanzee
Bonobo
Human
Gorilla
Old World Monkes

Common ancestor

30   20   10   0
Time in Millions of years

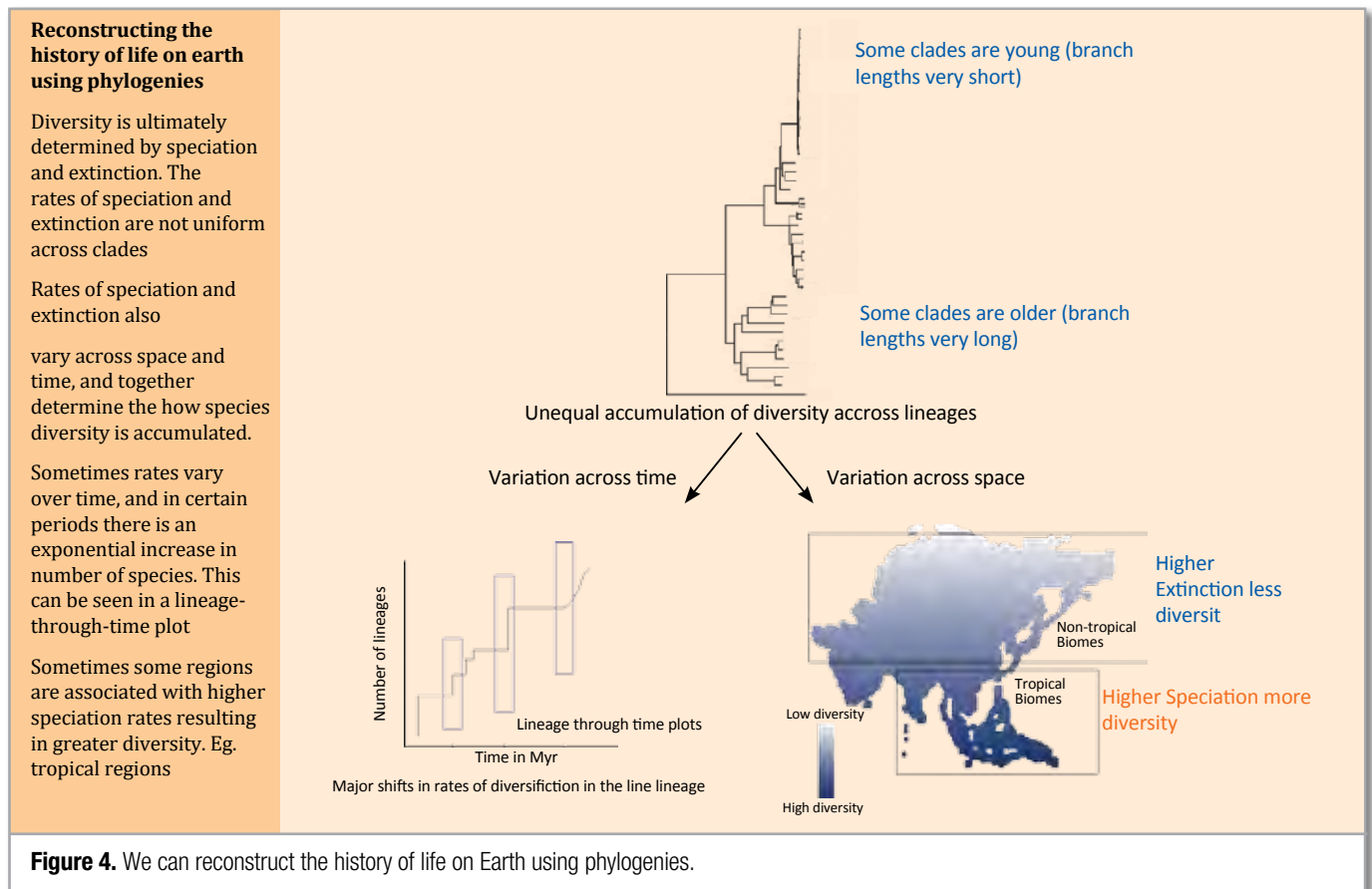**Figure 3.** How do we reconstruct relationships between primates?

and that some clades are much younger than others. Studying the spatial and temporal differences in the rates of diversification processes (speciation and extinction) helps us understand how biodiversity is built up. Although, we still haven't mapped the evolutionary history of the majority of life on earth, the explosion of genomic data has now given us much more power to detect and evaluate diversity. Instead of focusing on a few genes at a time, we can now look at almost the entire genome of a species. This gives us unprecedented access to the genetic history of species, allowing us to detect the historical factors that shape a species/individual across all time scales - past to present. While still in their infancy, a variety of methods are now being developed to generate phylogenies from such data.

What about patterns within species?

Stand at the bus depot anywhere in India and you can see a great diversity of people. India is just a microcosm for global diversity. There is enormous diversity within humans across the world - in terms of hair, skin colour, face structure and other morphological features. But, how diverse are we exactly?

Despite our very visible morphological diversity, humans show much lower genetic diversity between populations than chimps do! However, there are some clear geographic patterns in the distribution of genetic variations within human populations. **The study of how genetic variation within a species (and across closely related species) is partitioned across geographic space is the main focus of the field of phylogeography**. A comparison of this genetic diversity between different populations provides us with unique insights to the history of these populations.

For example, a comparison of the genetic diversity of the different human populations shows us that not only do African populations show the most genetic diversity, the genetic diversity of all other human populations are a subset of this African diversity. This has led us to conclude that anatomically modern humans (*Homo sapiens*) originated in Africa, and spread out from there to colonize the rest of the world. This historical movement (and all subsequent ones) are recorded in the patterns of geographic distribution of genetic diversity. And it's not just migrations - every single demographic event in our history, be it an increase or decrease in population, has left an imprint on our genetic diversity. As you can see, a lot can be said based on the partitioning of genetic diversity!

Reconstructing the history of life on earth using phylogenies

Diversity is ultimately determined by speciation and extinction. The rates of speciation and extinction are not uniform across clades

Rates of speciation and extinction also

vary across space and time, and together determine the how species diversity is accumulated.

Sometimes rates vary over time, and in certain periods there is an exponential increase in number of species. This can be seen in a lineage-through-time plot

Sometimes some regions are associated with higher speciation rates resulting in greater diversity. Eg. tropical regions

Some clades are young (branch lengths very short)

Some clades are older (branch lengths very long)

Unequal accumulation of diversity accross lineages

Variation across time

Variation across space

Number of lineages

Lineage through time plots

Time in Myr

Major shifts in rates of diversifiction in the line lineage

Non-tropical Biomes

Tropical Biomes

Low diversity

High diversity

Higher Extinction less diversit

Higher Speciation more diversity

Figure 4. We can reconstruct the history of life on Earth using phylogenies.

Gene flow: When individuals migrate between populations, they also carry any genetic variation that may be unique to their population to the new population. Thus migration results in the mixing of genetic variation, and prevents populations from becoming very different. This is therefore also called 'gene-flow'.

Alleles: Variant forms of a gene. Each gene occurs in two copies in every human. If the two copies are identical, then the person is homozygous, and if different heterozygous. Any individual can have only up to 2 alleles of a gene, while a population can have multiple alleles.

Mitochondrial DNA: The mitochondrion is an organelle that is found within the eukaryotic cell. These have their own small genome that codes for some proteins used within the mitochondrion. This DNA is called mitochondrial DNA, and is inherited only through women/mothers (as the sperm does not contain mitochondria when it fuses with the egg).

Cytochrome b: A gene that is located in the mitochondrial genome, and codes for a protein that is an important part of the oxidative pathway.
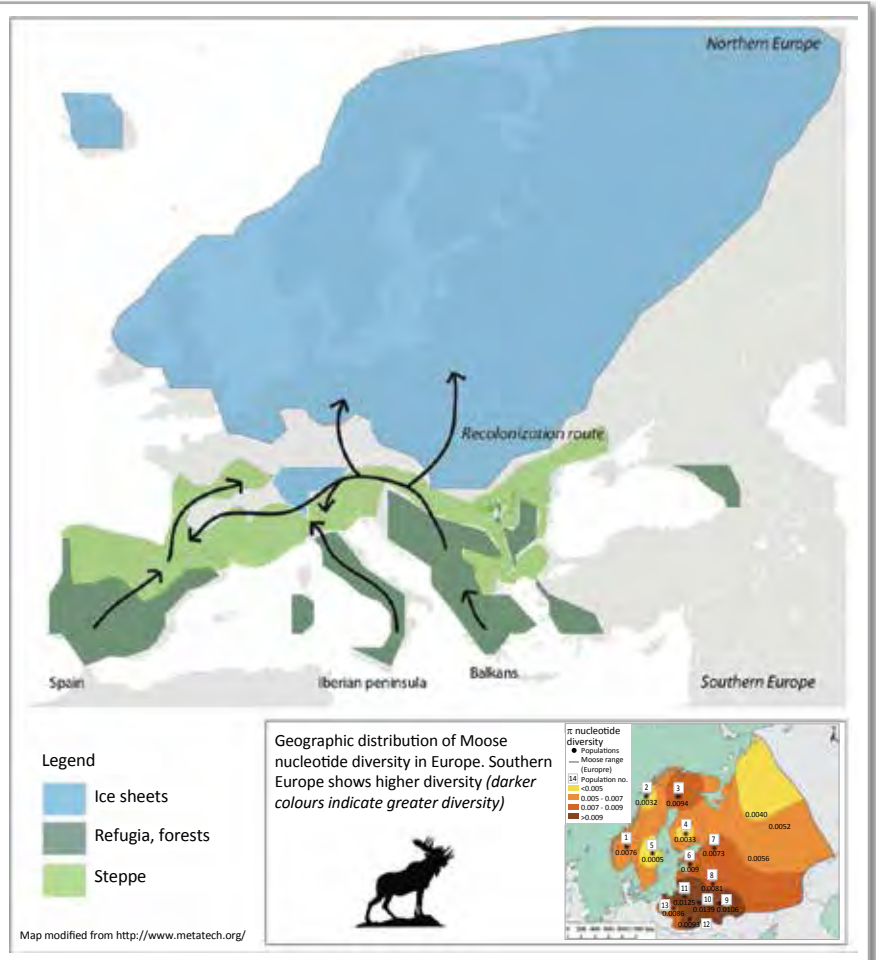
Single nucleotide polymorphisms (SNP's): Every difference in nucleotides between two individuals is an SNP. It is the most common type of difference between any two genomes.

In another example, a recent study attempted to estimate the genetic diversity of different tiger populations - in southern, western and central India. India has the largest number of wild tigers in the world, now restricted to pockets of forests across the country. Based on an analysis of shared and unique genetic diversity, scientists were able to quantify gene flow (as a proxy for migration or movement of individuals across populations) between the different Indian tiger populations, and determine if there was a change in the demographic histories of each population. This study indicated a decline in tiger populations as recently as 200 years ago (probably due to the practice of bounty hunting in the British Raj). Thus, just like a phylogeny, we can also infer the timing of demographic (population increase, crash etc.) events.

The confidence with which we can infer past events has been particularly aided by the development of the coalescent theory. This theory relates contemporary allele frequencies (distribution of alleles) in a population to the demographic history of the population, and has been applied to build 'gene-genealogies'. To give a simple analogy, these gene genealogies are family histories and animal pedigrees, only that these are at the scale of

Phylogeographic patterns in Europe

For the past 2.4 Mya, ice sheets have expanded and contracted at a regular rate over most of the temperate region, leading to global effects on distribution of biota. Particularly i Europe, ice sheets occupied most of the northern continent. This resulted in species being restricted in the non-glaciated regions in the south and subsequently recolonizing northern Europe when the ice-sheets melted. This was inferred by studying the genomes and the genetic diversity of populations across Europe, especially in Iberian peninsula, Spain and the Balkans showed higher genetic diversity than northern Europe. This suggests that southern Europe acted as glacial refugia for many of these species (Figure 3). This contraction and expansion in geogrpahic range and populationsize thus had consequences to the genetic diversity of the different populations of the species.

Geographic distribution of Moose nucleotide diversity in Europe. Southern Europe shows higher diversity *(darker colours indicate greater diversity)*

Legend
- Ice sheets
- Refugia, forests
- Steppe

Map modified from http://www.metatech.org/

genes and alleles. In many ways this is parallel to how phylogenetics works, except that gene-genealogies are limited to those allele frequencies that are influenced by population growth or decline. By mapping the history of the alleles in a population, one can understand, among other things, if populations have shown an increase or decrease in size over time.

Demographic history (population growth and decline) and migration patterns of species were initially studied through the geographic distribution of cytochrome b/mitochondrial DNA haplotypes (a set of DNA variations that tend to be inherited together). These early studies were based on one or few molecular markers targeting small portions of the genome. This limited clarity and the detail with which one could understand population subdivision and gene flow. Moreover, cytochrome b, being an important gene, also shows lower mutation rates. Subsequently, the use of microsatellite (nuclear) data, based on many more loci, offered more powerful ways to detect patterns. Microsatellites are regions of the genome that are largely characterized by repeated units of nucleotides. Owing to their genetic composition, they tend to mutate at much higher rates than cytochrome b or other genes. This allows them to rapidly accumulate mutations over shorter time periods, giving us a peek into more recent population and demographic patterns as well. Today, thanks to advances in genetic methods especially those relating to genomic data and analyses (including single nucleotide polymorphisms), we have more power to detect demographic history. With genomic data one can access much more of the genome, and obtain signatures of more recent events. Not just that – genetic data has become far more accessible because of decreased costs of sequencing as well as the advances in computing.

Conclusion

DNA is an excellent repository of history. Phylogenetics and phylogeography help us read and interpret this history of lineages and populations. In other words, they help us piece together the 'history' in the natural history. However, we cannot go back in time to observe our ancestors, or verify the phylogenetic trees that we build or the demographic histories we reconstruct. Thus, every phylogenetic tree and gene-genealogy is viewed as a hypothesis about the relationships between taxa and the patterns of branching. This is also why phylogenetic trees can change in the face of better or newer data.

Today, the genetic tools available to us have given us unprecedented access to, and an understanding of, the history of lineages, species and populations. DNA-based phylogenetic and phylogeographic approaches have revolutionized our understanding of the natural world and the processes that give rise to natural patterns. While there is a lot that we don't know yet, with better techniques, we can hope to learn more!

References

1. Gregory, R.T. (2008). Understanding evolutionary trees. *Evo Edu Outreach*, 1, 121-137

2. Rolland, J. (2014). Faster speciation and reduced extinction in the tropics contribute to the mammalian latitudinal diversity gradient. *Plos Biology*

3. Hickerson MJ et al. (2010). Phylogeography's past, present and future: 10 years after Avise, 2000. *Mol Phylogenet Evol*, 54(1): 291-301.

4. Hewitt, G. (2000). The genetic legacy of the quaternary ice ages. *Nature*, 405

5. Henn, B et al. (2015). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *PNAS*

6. Mondol, S et al. (2013). Demographic loss, genetic structure and conservation implications for Indian tigers. *Proceedings of the Royal Society B*.

7. Understanding Evolution. (2016). University of California Museum of Paleontology. 22 December 2015 <http://evolution.berkeley.edu/>

8. Maddison, D. R. and K.-S. Schulz (eds.). (2007). The Tree of Life Web Project. Internet address: http://tolweb.org

9. Baldauf, SL. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, 19(6), 345-51.

Krishnapriya Tamma completed her PhD National Centre for Biological Sciences, TIFR Bangalore. She is interested in large-scale patterns of species distributions, and the factors that influence them. Her PhD work focused on investigating biogeographic patterns of small mammals in the Himalayas. She can be contacted at priya.tamma@gmail.com.