An Introduction to Evolutionary Biology

Prof. Sutirth Dey

Biology Department, Population Biology Lab

Indian Institute of Science Education and Research (IISER) Pune

Week 10 Lecture 44

How new genes are born

Hi, so in our last discussion, we looked at how speciation happens primarily through the process of reproductive isolation. Now, if you remember our discussion from the first week, We said that Darwin said that the entire diversity we see around us. that has arisen because there have been continuous rounds of speciation, you know, over the last billion years. because of which, through descent with modification over time starting from some simple ancestors, The entire biodiversity around us has arisen. Now, if you think about it very closely, there is a bit of a missing link here. What is that thing? So, when we were discussing speciation, We saw in what way reproductive isolation sets in such that two populations that were previously able to exchange genes now they are not able to exchange genes anymore and thus they have become new species. But when they become new species like that, they are essentially becoming new species of the same type. In other words, when two salamander species or two bird species are becoming reproductively isolated from each other, They are becoming two different species of salamanders; they are becoming two different species of birds. They are not really converting it into something very, very different, you know they are not converting into let us say a snake or a mammal or something like that, right. In other words, for the mechanisms of speciation that we talked about, Those are talking about how more groups of the same type can arise. But if we have to give credence to Darwin's original claim that descent with modification has led to all of biodiversity, Then we have to have a way to figure out how exactly biological organisms lead to new traits or, rather, generate new traits. Until and unless you explain how biological novelty originates, it is very hard to see how you can go from what you know. Simple organisms evolve into more and more complicated organisms, or, you know, things like the peacock's tail evolving, etc.

All that one cannot even think about unless one understands how biological novelty originates. And that is something we are going to talk about in today's discussion. As I am telling you here, this treatment is primarily after Futuyma and Kirkpatrick. However, there are one or two extra things here and there that you know I might be able to say. Now, this particular question is: where does novelty come from? Of course, you know Darwin thought about it; many other people thought about it.

And one of the things that they realized is that, in many cases, what shows up as a trait today It is essentially some repurposed organ that arose for some other purpose. So, for example, you know, think about us humans; we have these opposable thumbs, right? And because of that, We are able to grip things, and this is supposed to be very, very important in the evolution of humans. Because this ability to grip is what allowed us to make and use tools. Now it turns out we are not the only organisms that are able to grip things. Another famous example of things that can grip is the giant pandas.

And you can see, you know how it is able to grip this bamboo and eat that. But if you look at the panda's thumb and compare it with the human thumb, you will find a crucial difference. What is it? So, in humans, the thumb, which is this one, is the fifth digit in the hand. Now, in the context of the pandas, if you look at them, they will have all five of their digits. So the fifth digit is the one that you know is homologous to our thumb, but it is not the opposable one.

The opposable one is a sixth digit over here, and that is not really a digit. That is essentially one of the wrist bones that has become modified to serve that purpose. In other words, in many cases when organisms are trying to evolve a new trait, All that they do is take an existing thing and somehow modify it. Change it in some way so that it can be used in a new way. Now this repurposing is not only something that happens at the level of the organism.

It can also happen at the levels of genes. So I will give you a very nice example. So, for example, what I am showing you is a schematic of the vertebrate eye. And you know that in the vertebrate eye, we have this lens over here, and this lens is very important because. It gathers the light and then focuses it on the retina, from which, through nerve conduction, the signal goes to our brains.

Now, obviously, in order for the lens to be able to do this, it has to be a transparent object. If it is not transparent, light will not pass through it. So in vertebrate lenses, the main protein that takes care of this function, This is the transparent set of proteins known as the crystallins. Now it turns out that different vertebrates have different combinations of crystallin proteins in their lenses. So all of them, the eye is doing the same job; the lens is doing the same job of letting the light pass.

But the proteins that are combining to form those lenses are different. And just to show you how different they are, So here you know are a number of proteins, crystalline proteins in various kinds of organisms going all the way. From frogs to, you know, humans to rabbits, but what is more important than this is this list over here. So these are the lists of the ancestral proteins and their functions, which have finally led to the clear crystalline proteins in the eyes. So here you find that a large fraction of them is actually heat shock proteins.

So these are proteins that are supposed to be stress-resisting proteins, as the name suggests. Typically related to heat shock, you know, withstanding heat shock, but also playing other kinds of roles. Then there are others, you know; for example, this is a reductase, an NADPH-dependent reductase. Then you have other kinds of enzymes; you know this is a dehydrogenase, there is a deaminase, and so on and so forth. I mean one of them is the alcohol dehydrogenase gene, you know, the one, for example which forms the crystalline protein in the camels, right? This one over here, zeta (ζ). So, this tool tells you that the crystalline protein has actually arisen in different organisms at different points in time. Which is what is shown by that phylogeny. And in each case, each lineage has actually taken a very different protein. and has ended up repurposing that for the purpose of forming the lens.

So, this tells us that proteins do change their function, and they can change their function. Now, there are two ways in which a protein can change its function. What are the two ways? First, it can change its own structure, the amino acid structure. which obviously requires changing the sequence of the protein-coding gene itself. And the second thing is, instead of changing the structure, It can also change its expression pattern, how much is being expressed, where it is being expressed, and so on.

Of course, a third thing is that there can also be, you know, post-translational modifications that can also. you know lead to change in its function. Now, if a protein ends up changing its structure or its expression pattern, what is going to happen to its original function? Because after all, if the protein is there, it has to be playing a role, right? So, if suddenly a heat shock protein, a stress-resistant protein, starts forming in the eye lens, What is going to happen to the heat shock function? So, there has to be a way by which the original function can be retained in the organism. And then, on top of that, the organism, you know, the protein should be able to take on a new function. So, how exactly can this happen? Now, it turns out that there are many ways in which this can happen, and the first and simplest way is to form a new gene.

So, where exactly, you know, how will a new gene be formed? Where will it come from? So, it turns out that there are four major ways of gene formation. And these four major ways are what we are going to cover in our discussion today. And these are as the list says: gene duplication, exon shuffling, de novo gene birth, and horizontal gene transfer. So, we will start with gene duplication, followed by mutations. So, what exactly is gene duplication? So, as the name suggests, there is a duplication of an existing gene or a large chromosomal segment containing a gene.

So, two or more genes that have arisen due to duplication are known as paralogs.

Whereas, if you have one gene that has mutated over time due to mutations or other factors. And a new, you know, form of the gene is there. So, these two are related directly by ancestry; these are known as orthologs. So, what is the logic behind gene duplication? How does it solve the problem? So, the whole idea is that if you have multiple copies of a gene, then one copy will continue to perform the original function whereas the other copy is now, in some sense, freed up and can mutate, taking on other functions and so on. So, it turns out that gene duplication plays a very, very key role in evolution. And about 1% of human genes have been duplicated every million years. Now, think about it: that is, you know, although a million years sounds very big for us, evolutionarily speaking, it is just a blink. So, 1 percent of human genes have been duplicated every million years, and, About 1,400 duplications have been fixed in humans and chimpanzees since we diverged about 7 million years ago.

So, this tells you how critical this is. Now, what are the ways in which genes can be duplicated? So, there are three or four ways in which it can happen. The simplest way is known as unequal crossing over. So, when you have meiosis during that time, you know the gene segments are supposed to be exchanged. But instead of exchanging, what happens is that there is a duplication in one chromosome and the other chromosome this thing does not come and you know it just there is a deletion over there. So, in this way, if the part that is getting duplicated contains a gene Then there is a gene duplication that is occurring on one of the chromosomes. The other way in which it can happen is known as replication slippage. So, during DNA replication, sometimes the DNA polymerase gets dislodged from the DNA and then it sort of comes back. But in the process of doing that, sometimes it ends up copying the same sequence more than once.

This is particularly common in those regions of the genome where there are repetitive sequences. The third way in which it can happen is retrotransposition. So suppose you have a gene and the gene has been transcribed into an mRNA, but then after that, instead of going into The translation process involves the transcribed mRNA being reverse transcribed into complementary DNA or cDNA. And then this cDNA, you know, comes back and gets reintegrated into the genome at some random location. So, this is what is known as retrotransposition, and obviously, this can lead to the formation of new genes.

Now note that since this process is occurring with the mRNA, the introns are cut out of the mRNA. So the cDNA typically does not have the introns, and therefore a simple way for people to recognize That a gene has arisen due to the retrotransposition of a different gene, obviously there will be sequence homology. But at the same time, the intron portions will be missing in the new gene. So that is how you know it is a retrotransposon gene. And of course, the other way in which you can have gene duplication is if the entire genome ends up duplicating itself.

So you know when it is happening within the same species, that is what is known as autopolyploidy, and, As we have discussed, this is something very common in plants. Now let us look at some examples of gene duplication. So this is a very beautiful-looking langur. This is the red-shanked douc. It is found in Southeast Asia, and it turns out that, unlike most primates, this one actually eats leaves.

Now you know that leaves contain cellulose, and therefore, leaves are very, very hard to digest. So the other group of organisms that can digest leaves is the herbivores. Many of them have these four-chambered stomachs and they have bacteria. You know symbiotic bacteria that live inside the gut that help them in digestion. So essentially what happens is that this bacteria ends up digesting the leaves.

And then the organism, the host, ends up digesting the bacteria, and that is how it gets the nutrition. Now it turns out that in this particular Duke Langur, they use one of the enzymes known as RNASE1B. And this RNASE1B digestive enzyme rose by gene duplication about 4 million years ago. And the gene from which it arose is the pancreatic ribonuclease gene, RNASE1. So, the RNASE1 gene is still there inside the douc, but its function is entirely non-digestive.

So, after the RNASE1B gene was created, it actually acquired lots of mutations very, very fast. So, there were 9 amino acids that ended up changing, and because of these changes, the RNASE1B became a digestive enzyme. And more critically, now it was able to function in the small intestine. Where it had, you know, it could function in a slightly more acidic environment. So, these mutations enabled this thing to function in an acidic environment which was really good for these guys, the doucs, to be able to, you know, become leaf eaters. So, this is how a new function is being gained. Now, another very, very important example of gene duplication is in terms of the vertebrate sodium channel. So, you know that in our nerves, we have these voltage-gated sodium channels. These are very, very important for the purpose of nerve firing.

Now, if you look at the structure of these genes, you will find that they are composed of four domains. Each domain, in turn, is composed of six units: six transmembrane segments. Now, if you look at the sequences of these transmembrane domains, I am sorry; these transmembrane proteins or segments are in each domain. You find that you know the 6 in domain 1; you can call them 1, 2, 3, 4, 5, 6. The sequences are extremely similar to the six in the, you know, second domain, the third domain, and the fourth domain which basically means that the first segment of domain 1 has the same sequence as the first segment of domain 2. As the first segment of domain 3, and so on and so forth for all the 6 segments. So, this suggests that there is probably some relationship here, and people have actually figured it out. That this entire thing happened because there was a duplication, or rather two duplications, one after the other. So, the first one became two, and then each one of them gave rise to two more; that is how you ended up getting four and so, from one locus, you ended up getting this entire thing. Now this is an example of gene duplication. And now we are going to talk about the special case of gene duplication, which is retrotransposition. So in human beings, we have this gene known as phosphoglycerate kinase 1. PGK1, and this is a very important gene because it is involved in glycolysis, you know.

So it is found on the X chromosome, contains multiple introns, and is expressed throughout the body. So it is a very important gene. Now it turns out that human beings also have another gene called the PGK2 gene, and this is found on the 6th chromosome. And very interestingly, when you look at the function of the PGK1, I am sorry, the sequence of the PGK1 and the PGK2 genes. You will find that there is a huge amount of similarity between them, except for the fact that the PGK2 gene does not have any of the introns that are present in the PGK1 gene. So, this clearly tells you that this gene, PGK2, has arisen due to retrotransposition from PGK1. which has moved it from the X chromosome to the sixth chromosome. And it turns out that this PGK2 is now under the control of a different promoter.

Now, this is very important. See, retrotransposition is simply going to take the cDNA and put it somewhere else. Randomly, somewhere else in some other, you know, place, or sometimes in the same chromosome, it does not matter. But wherever it is going, it has to be reasonably close to a promoter region or within the influence of a promoter region. without which the gene will not be expressed. So, this is the problem with retrotransposition: until and unless you know it gets integrated in an opportune place where it is under the control of, you know, different controlling elements; without that, the gene is not going to work. And in this particular case, it turns out that the PGK2 gene did find itself in the vicinity of a proper promoter, and now, Today, this gene only expresses in the testes of humans. Because of that, it actually ends up playing a big role in sperm motility in humans. Now, fine, we saw how genes duplicate; we saw examples of genes duplicating. Now the question is what exactly happens when a gene is duplicated.

Now there are two major possibilities. One possibility is that, as I said, the duplicate gene is not even functional to begin with. This can happen in many ways. It can happen because, suppose the position you know, the part that got duplicated does not contain the entire gene. So, it is just a fragment that does not work, or you know, as I said, particularly in the case of transpositions. It does not have the regulatory elements that are needed to express the gene properly; in which case, it will not function or, you know, in many cases, the post-translational, you know, transcriptional splicing regulation is very, very important. But suppose the duplicate has been formed due to transposition, in which case the introns are not there, which means that All the post-transcriptional splicing regulation will not happen, which in turn might mean that the gene may not function. So, in all these cases, we are going to get a duplication event. We are going to get some fragments, either full or, you know, incomplete, but that fragment will not lead to any product. And in these cases, such non-functional genes are known as pseudogenes.

The point with the pseudogene is that, because it is non-functional, it is essentially selectively neutral, which means that it will end up. Gathering mutations in the genome will keep on making it more and more unlike what it rose from. So, basically, selection will not work on this. On the other hand, the second possibility is that you actually have a functioning gene. But if you have a functional gene, then there are four possible fates of a duplicated functional gene.

What are those? So the first one is that suppose the gene is functional, but it is not really providing any fitness benefits. In other words, whatever the original gene was producing, the quantum of that is sufficient. And therefore, this new gene in the organism does not even make it feel whether it is there or not. Now, if that happens, then there is a possibility that this might get lost because of any bad mutation, say deletion or some kind of. You know other kinds of things that lead to loss of function mutations; none of those are going to be seen by selection.

And therefore, at some point, there is a possibility that this might become an inactive gene. In which case, it will again become a pseudogene. The second possibility is that the duplicates actually retain their original function. This retention is something that is selectively favored. When can this happen? This can happen if the amount of the gene product that is needed is somewhat large.

Having an extra copy of the gene allows for an extra amount of protein in the body. That particular protein, whatever is coded by that gene, is beneficial for the organism. So, if such a thing is happening where an increased quantity of the protein is useful, Then there is another possibility that there might be subsequent duplication events, because of which. The number of repeats in the genome is actually the number of duplicate genes for that particular organism. You know the original gene keeps on increasing, and it actually varies between individuals in the population.

So, this can lead to variations in the number of copies of the gene that various individuals in the population carry. And this phenomenon is what is known as copy number variation or CNVs. And as I said, CNVs can affect the gene dosage, so that the level of the product of the gene in the body of the organism is altered. And in that sense, it can become a very significant source of genetic variation.

Now this was number two: CNV. The third thing that can happen is what is known as new functionalization, which is that the redundant gene copy can now Start accumulating mutations, and in that process, it might acquire a new function, thereby leading to some kind of novelty. And we have already looked at one such example: the RNASE1B in the red-shanked Douc Langurs. We will look at one or two more examples as we go along. And the fourth thing that can happen is known as subfunctionalization. What is that? Now suppose the original gene was performing more than one function.

And when there are two copies, one copy takes care of one part of the original function. Other copy takes care of another part of the original function, such that it was the job. That was earlier done by one generalist gene; now there are two specialist genes doing two parts of the job. So, this is what is known as sub-functionalization. So, I will give you examples of all these things: copy number variation, neofunctionalization, and subfunctionalization.

Now, sometimes, though not always, the two can even happen hand in hand. Neofunctionalization and sub-functionalization, and I am going to talk about them too. So, let us talk about an example of copy number variation. Now, you know that we humans have this enzyme called alpha amylase in our saliva. And this is very important because it allows us to break down starch. So, starch digestion starts right in the mouth, which is why if you take some bread and you chew it for slightly longer the entire thing starts you know tasting sweet.

The starch has started breaking down into its component glucose and other substances. Now, this particular enzyme, alpha-amylase, is coded by a gene called AMY1, which shows copy number variation in humans. Now it is known that if you have more copies of AMY1, then you have more alpha amylase in your saliva. And if you have more alpha amylase, then that means you are going to become better at starch digestion.

Now this line of reasoning is leading to a prediction. What is the prediction? The prediction is that you know different populations have different levels of starch that they eat. Some people or some populations of individuals eat a lot of meat, let us say. And there are other places where individuals eat lots of starch, and you know everything in between. So, the prediction here is that populations whose diet is more meat-based do not really need the alpha amylase thing. So there will be the number of copies of alpha amylase; sorry, it should be the other way around.

Populations whose diet is more meat-based should have a lesser number of copies of AMY1. Sorry, whatever is written on the slide is wrong; just correct that. So populations whose diet is more meat-based should have a smaller number of copies of AMY1. Compared to those whose diet has more starch.

So, do we have any evidence to believe this? Turns out that we do. So these people, Perry et al., what they did was look at different populations, you know. where the amount of meat that they eat versus the amount of starch that they eat is very different. So here I am showing you an example from one individual from the Biaka tribe of Africa.

So these are hunter-gatherers, and their diet is highly meat-based. They are being compared to an individual who is Japanese from Japanese society. Which, as you know, is a more modern society, and there, of course, they do eat meat; they do eat fish, but The amount of starch that they eat is much larger than, for example, that of the Biaka. So, what Perry et al. did was extract, you know, tissue from both the Biaka and the Japanese and they ended up, you know, looking at their chromosomes by coloring them red and green. So, each red-green thing that you see over here is showing you the AMY1 locus in that individual. So, you can simply count the Biaka individuals; this has 3 copies on one chromosome and 3 copies on the other chromosome. So, there are a total of 6 copies of the AMY gene in this individual. Whereas, if you look at the Japanese individual on one chromosome, this person has 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

And the other chromosome that person has is 4. So, this person is carrying 14 copies of the AMY1 locus. So, this clearly shows you how there can be huge variation in terms of the copy numbers. However, whatever I showed you over here, this is just one individual from, you know. The meat-eating population compared to one individual from the starch-eating population. But the copy number variation definition that I gave you talks about this stuff at the population level.

So, obviously, in order to understand this properly, you need to look at the population level. So, it turns out that they did the same thing; the same paper, you know, looked at this problem. So, here what they did was take 7 populations; 3 of those populations incorporated diets that had a lot of starch. whereas four of the populations had traditional diets that had either little or no starch. So, they call the latter the low-starch and the former the high-starch diets or high-starch populations.

And here on the x-axis, they have the AMY1 diploid gene copy number. So, over both chromosomes, how many copies does each individual have in a histogram of that? And what proportion of individuals have that many copy numbers? That is what is on the y-axis. It is a normal histogram. And this orange thing shows the low starch population, while the blue thing shows the high starch population. So, you can see that the proportions of the low starch populations are larger in the low copy number case. whereas the moment you go somewhere above 6, Typically, it is the high-starch populations that have a greater number of, you know, greater copy number.

Basically, I am telling you that on average, the high starch populations have a much greater copy number compared to the low starch populations. And how much greater was greater? So, on average, the populations with a high starch diet have 6.

72 copies of AMY1. While those that have a lower starch diet have about 5.44 copies. Great. So, this is an example of copy number variation. Now, we will look at an example of neofunctionalization. So, we have already looked at one example: RNASE1B. We will look at an even starker example, that of color vision in some primates.

So, in our eyes, we have a gene that is, or rather, we have many multiple genes. Known as opsin genes, these lead to the formation of opsin proteins. Now, these opsin proteins lead to the formation of different colored pigments. which have their maximum absorption in certain you know wavelengths. Now, the way we see color is that all primates see colors, and all vertebrates see colors through the action of these pigments.

And the more pigments you have, the more colors you will be able to see. So, if you look at ancestral primates, they had two opsin genes. One was on chromosome 7 and autosomal, and the other was X-linked. So, the gene that was on the autosome, on chromosome 7, the autosomal one, The protein coded by that was able to absorb light at the short wavelength. So, the blue side wavelength, whereas the other one, which was on the X chromosome, That one coded for a protein that was typically absorbed in the longer wavelength, the reddish side. So, because of this, ancestral primates had dichromatic vision, which means that they could only see those colors.

This could be created by mixing these two pigments or by the effects of these two pigments. Therefore, the color vision in these primates was relatively limited. However, at some point in the ancestors of apes and the Old World primates, who are the Old World primates? The baboons, the langurs, the macaques, these are the old world primates. So, at some point in their ancestry, there was a gene duplication that occurred, and this happened around 35 million years ago.

And this happened for the gene that was on the X chromosome. When that happened, it was a tandem duplication. Which means that the new copy was placed right next to the old copy on the X chromosome. Now, when this happened, then one copy, it actually retained its original function, And this is what today is known as the long-wavelength opsin. And this is one that absorbs light maximally at, you know, 560 nanometers. The second copy actually started gathering mutations, and people have even shown what those mutations are. There are three mutations that happen one after the other, and because of these three mutations, The optimal wavelength that it could absorb went down to 530 nanometers, not much, just 30 nanometers; that is it.

But now, because of this reduction, it was able to detect the medium wavelengths. And therefore, suddenly, from a dioptic vision, the apes and the old world primates were able to now have a trioptic. Sorry, trichromatic vision, because of which they were able to see many more colors. And this, you know, actually is supposed to have been a major evolutionary boost to these two lineages. The apes and the Old World primates, because when you see more colors, Obviously, you are able to differentiate the world much, much better. So, one way in which people have hypothesized that this might have played a huge role is in terms of their ability to distinguish a ripe fruit from an unripe fruit against a background of foliage, the green background of foliage. So, the old world primates with their trichromatic vision are supposed to be much, much better. In terms of doing that as opposed to the new world primates. And this is supposed to be an evolutionary advantage that evolved because of the neofunctionalization of that gene on the X chromosome. Now, here's a quick example of sub-functionalization. So, remember that sub-functionalization is where the original gene has a broader function and then there is a gene duplication and the function of that now becomes, you know, the original general function now becomes partitioned into two or multiple genes. So, this example comes from the zebrafish, a very popular model system, Danio rerio. Now, in zebrafish, there is a certain gene called the engrailed gene. whose product helps both in the formation of muscles and in the formation of the nerves in the retina.

Now, it turns out that in these fish, there are two such genes, engrailed genes, so-called eng1a and eng1b. And these two genes were formed due to a duplication event in the teleosts, the ray-finned fishes in that lineage. Now, it turns out that the ancestral Eng1, which Eng1 is still found in mammals and birds. In both these organisms, Eng1 is expressed both in the head and in the muscles, right? But when it comes to the zebrafish, eng1b expresses only in the hindbrain and in the spinal neurons whereas Eng1a is expressing itself in the pectoral appendage bud. So, basically, over here. So, which suggests that the earlier function of getting expressed all over the body is what the ancestral gene was doing. And now, when the sub-functionalization has happened, One copy of the gene is expressed in one place, while the other copy of the gene is expressed in another place. So, this is a spatial sub-functionalization that has ended up happening. Now, it turns out that every mechanism that I showed you, all of that, It is not that these are happening in a vacuum; it is not that they are happening one at a time.

Sometimes these things actually end up happening together, leading to a very fascinating array of, you know, molecules. And one such example that I am going to share with you is the globin gene family in primates. Now, you must have heard of hemoglobin, the molecule that allows us to carry oxygen and carbon dioxide in our blood. You might have also heard of myoglobin, right? The molecule that stores oxygen in our muscles. Now, it turns out that hemoglobin and myoglobin actually come from one ancestral gene and that divergence happened about 600 to 800 million years ago. So, that is not what is shown here. Now, please note that whatever I am showing you here is not a phylogenetic tree. This is just a flowchart showing how the divergences happen. But I have placed the values in some locations. So, after this divergence occurred between hemoglobin and myoglobin, The protoglobin gene, at some point, underwent a duplication.

And then, after the duplication, there was a lot of divergence that occurred due to mutation accumulation. But they were still on the same chromosome. So, at some point, you knew you had two gene copies that had become sufficiently diverged, One led to the alpha-globin family; the other led to the beta-globin family. So, the alpha globin family is this one and the beta globin family. The ancestor of the beta-globin family is this one; this is the ancestor of the alpha-globin family.

Then, at some point, there was a transposition that occurred. Now, because of that transposition, the progenitor of the alpha globin gene that came from the 16th chromosome, Whereas the progenitor of the beta globin came to the 11th chromosome. Now, where the ancestral one was, I could not figure it out. But what we know today is that one set is on chromosome 16; the other set is on chromosome 11. Now, about 300 million years ago, the progenitor of the alpha globin family divided into two parts.

One part became the so-called zeta, you know, family, and the other went towards the alpha family. Then, about 40-50 million years ago, there were again more duplications. And three such genes were formed: alpha (α)1, alpha (α)2, and one more gene psi-alpha (Ψα)1, which was formed. But it essentially died after that; it became a pseudogene. Similarly, at this end, the zeta gave rise to what is known as the Z globin gene.

It gave rise to another one (Ψζ), but that one also died. So, together all of these form what is known as the alpha globin family. On the other side, what happened was that about 150 to 200 million years ago, there were, you know, three duplication events. So, you have the progenitor of the epsilon, which gave rise to the epsilon (ϵ) globin. Then you have the progenitor of the gamma globins. And then, at some point about 100 to 140 million years ago, that again divided by duplication to lead to two different things: the G-gamma (Gγ) and the A (Aγ) gamma forms. And then the third thing that had arisen was the beta ancestor, the ancestor of the beta thing, and that beta ancestor. about 100 to 140 million years ago, again ended up duplicating and diverging, and forming three different genes. Today, two of those are still alive, two of those genes are still functional; one is called delta (δ), and the other is called beta (β). And one of them actually ended up dying, and that is the pseudo, you know, psi-beta (Ψβ) pseudogene.

And all these together form the beta globin family. So, you can see that this entire thing has happened due to duplication, divergence, and non-functionalization. But this stuff that has happened, the consequences of it are amazing, and now we will discuss the consequences. So, as I said, the hemoglobin and myoglobin diverged about 600 to 800 million years ago. It looks like that happened because there was a whole genome duplication event.

Now, because of this, there was a very nice physiological division of labor that ended up happening. Why? Because the globin family is able to attach to oxygen, However, the way it attaches is going to determine what the function will be. So, for example, if you want a molecule to store oxygen, then you need it to bind to oxygen somewhat strongly. So, that is what happened to myoglobin. Whereas, if you want something to act as transport, That is something that should be able to hold it, but not so tightly that it will be unable to give it up.

So, it has to hold it in an optimal way such that, under certain circumstances, it will take the oxygen. In certain other circumstances, it will give out the oxygen, right? And that is what hemoglobin specializes in. But just one hemoglobin molecule, or rather one globin molecule, would not have been able to do it. So, what happened is that the proto-hemoglobin chain duplicated again.

And giving rise to the progenitors of the alpha and beta globin genes, as I showed you. This happened about 450 million years ago. And after that, there was a lot of functional divergence due to duplication and other factors. Because of this, we finally have so many alpha family and beta family genes. Now, it turns out that the hemoglobin we have today is a tetrameric molecule. There are two molecules from the alpha family and two molecules from the beta family.

And that is why it has an alpha (α)2 beta (β)2 configuration. And this configuration of having four different globin proteins coming together and forming one big, huge protein, This is what has allowed hemoglobin to become so effective in terms of transporting oxygen. And this would not have been possible if you had known. All these different globin forms had not evolved through all those genomes; I am sorry, gene duplication, and so on.

But this is not all. There is more to the story. What is it? Remember I told you about a form called gamma (γ) globin. Now, it turns out that subsequent duplications within the beta globin gene family resulted in multiple things. Including the gamma (γ) globin gene.

Now, this gamma (γ) globin, these genes, or the proteins that they are coding, These proteins are actually part of the fetal hemoglobin. So, in mammals, and particularly in primates, The fetal hemoglobin has very different properties compared to the adult hemoglobins. Why is that so? That is because, think about it, the fetal hemoglobin needs to be able to extract oxygen from the mother's blood, right? So, if the fetal hemoglobin has hemoglobin or if you know whether the fetus is carrying hemoglobin that has the Same properties as the mother's hemoglobin; then it will not be able to take the oxygen from the mom's blood.

That is why fetal hemoglobin needs to have a greater oxygen binding capacity. which it accomplishes by combining two alpha (α) units with two gamma (γ) units. This innovation is what allowed mammals to evolve. Only then can the formation of the placenta and the fetus extract oxygen from the mother's blood supply, evolution could occur. But that is not the whole story yet. So, in most vertebrates, the same hemoglobin ends up transporting oxygen throughout its life. But as I showed you, in mammals, it is fetal hemoglobin that is transporting oxygen during the fetal stage. Now, this is excellent as long as the fetus is inside the mom's body, but at some point, the fetus comes out; you know, it is born. When it is born, if the fetal hemoglobin continues, then that is going to be a problem for the independent offspring. Why? Because the independent offspring has to have its own oxygen supply, right? And at that point, if the fetal hemoglobin has a higher affinity for oxygen, then that will be an issue.

Therefore, a few months after birth, fetal hemoglobin is no longer made. And that is when the entire oxygen transport shifts to adult hemoglobin, hemoglobin A, which is alpha (α)2 beta (β)2. And this shift, again, is very, very crucial for independent life. So, after the evolution of the gamma subunit through duplication, There has been a temporal subfunctionalization between the beta and gamma subunits. In other words, earlier the alpha (α)2 beta (β)2, HbA was able to take care of.

The entire concept of the life of the entire oxygen supply throughout its existence. But now, because of the mammalian way of life, you need the gamma thing: the fetal hemoglobin. But you need fetal hemoglobin only for one part of life; for the other part of life, you need the other one. So, there has been temporal sub-functionalization between the beta and gamma subunits.

In terms of forming fetal hemoglobin and adult hemoglobin. And this entire thing was crucial for the evolution of life and birth in mammals. If this had not happened, then the evolutionary innovation of live birth would not have occurred. The placenta and the fetus getting all the nutrition from the mom, etcetera, etcetera; none of this would have been possible. So, this shows you how all these you know gene duplication, neofunctionalization, subfunctionalization, how all these things are actually interacting with each other in a rather complicated dance, so to speak.

In order to lead to novel functions in biological systems. Now, we started by asking how one creates new traits. And we talked about the formation of new genes, and then we were done with gene duplication. Now, we are going to look at another way of doing the same thing, which is exon shuffling. So, what is exon shuffling? So, exon shuffling is when you know what exons are, right? So, you know that in many organisms, the genes do not come in one continuous ATGC sequence.

So, you have, I am sorry, I will take that back. They do come in a continuous ATGC sequence. But the final thing that is going to be translated does not come in one continuous stretch on the genome. So, what happens is you have stretches that form, you know, the final product, which has information about the final protein. That is going to be coded, and these are known as exons; between the exons, you have stretches of DNA. which, after the mRNA is transcribed, are cut out and are known as introns.

So, two or more exons, the protein-coding regions from different genes, can be brought together ectopically, meaning from the outside. The same exon can be duplicated to create a new exon-intron structure, and this process is what is known as exon shuffling. So, just to give you a visual, you know, feeling for this, let us have these two genes. Gene 1, which has three exons—1, 2, and 3—and gene 2, which has three exons as well, let us call them A, B, and C. So, let us assume that what happens is that you know this part, B and C.

This comes, and this gets shuffled into this intron between 1, 2, and 3, and this three-part structure now comes over here. So, what you will get is exon 1, intron, exon 2, and then this intron is modified. Because these two exons go and sit there, the third thing that you know is that this exon 3 from here actually shifts to this one. So, what I am showing you here is a symmetric case, but it does not really need to be symmetric. It is entirely possible that you know, let us say, B and C from here; just come and sit over here, and then you have 3 over here.

So, there does not need to be a reciprocal exchange; that is all I am trying to tell you. Now, how does this happen? Retrotransposition is the simplest way in which it can occur. There is also something known as illegitimate recombination; things that should not recombine end up recombining. I would not go into the details; you can figure out the mechanics on your own. And the resulting protein that ends up being a mosaic of many different domains.

And sometimes it is just a hodgepodge that goes nowhere, but sometimes it can actually potentially lead to a novel function. So, I will give you one, and actually, I will give you two examples of this. So, in humans and in many other organisms, we have something known as tissue plasminogen activator. It is an enzyme that breaks down blood clots, and because of that, it is very, very important. Because if it is not there, then obviously the chances of your getting a stroke or a heart attack are increased severalfold which is why it is a very, very important enzyme. It is used tremendously for treating strokes worldwide. Now, if you look at the structure of this enzyme, you will see multiple domains. So there is one domain; this is, you know, a protease domain. This is where the cutting actually happens.

But there are other domains, you know, for example, these two Kringle domains. This EGF domain, this finger domain, and so on and so forth. Now what people have done is taken the sequence of each domain separately. And they have tried to map it with sequences of other, you know, genes for other proteins. By doing so, they have actually figured out that the exons of these genes have all originated in different places. So, for example, this Finger module that I am talking about here has come from a gene known as the fibronectin gene.

Similarly, this protease module over here has actually come from the trypsin gene. And this EGF module that I am talking about is essentially, you know, the epidermal growth factor. If you are studying cancer, you will come to know about it. But you know you can see it is a growth factor thing. So, essentially, this entire gene is a hodgepodge of exons from different places all coming together, and, Forming a novel gene with a novel function that is nevertheless extremely important for the organism.

The second example that I am going to give you for exon shuffling is a locus called the Jingwei locus. So this locus is found only in Drosophila teissieri and Drosophila yakuba, two species of Drosophila. And the way this locus is formed is very interesting. So this is the Jingwei locus. So in the Jingwei locus, if you look at it closely, you will realize that the first three exons, These three exons are actually just duplicates of another gene known as the Yellow Emperor gene, the Ymp gene, right? So these three exons, along with the two introns in the middle, are these.

Now, this is one duplication already. Now what has happened is this: this is Drosophila, right? So these live on they live on rotting fruits etc. Therefore, their ability to tolerate alcohol is a very big deal because rotting fruits are typically high in alcohol. And therefore, if you cannot tolerate alcohol and you are a fly, then you are in trouble. So, in order to break down alcohol, they have this thing known as the Alcohol dehydrogenase Adh gene. So what has happened is that after these three exons, there is a large intron in the Ymp gene.

So this entire thing is the entire Adh gene minus the exons, but along with the stop codon. This entire thing has actually gone, and it has now started sitting over here right in the middle of this intron. Now, how can this happen? The only way wherein you have a duplication, but a duplication only of the exons and not of the introns, is through retrotransposition. Therefore, the guess is that this entire thing is actually the retrotransposed, you know, ADH gene, and you know, looking at. How many mutations have accumulated? People have figured out that this happened about 2 million years ago.

So, and this is the beauty of this entire thing that has come, and also the stop codon. So if you look at the original Ymp gene, obviously, you know there is a lot more after that. But because this stop codon has now become inserted here, now this thing, I mean, it already had a promoter in this direction, right for the Yellow-emperor gene. So this thing is now essentially behaving like a separate and independent gene, and this is what is known as the jingwei gene. So it comes from a Chinese myth in which, you know, a princess had metamorphosed into a new form which is pretty much what this gene does. But the interesting bit is that this gene is active; it is a functional gene, and it is seen that it is expressed in the testes. It is thought to be involved in the metabolism of hormones and pheromones. Which, obviously, you know if you are a fly; these are very, very important things for your physiology. So this is an example of exon shuffling. Now we are going to get back, and we are going to look at the third way, which is de novo gene birth.

So, as the name suggests, this is not coming from existing genes; this is essentially, you know, non-coding gene sequences. You know that in the genomes of many organisms, there are massive DNA sequences that do nothing. So some of those "doing nothing" DNA sequences can actually lead to the formation of new genes. So, yeah, a non-coding sequence leading to a new gene. Now, how exactly will this work? In order for this to work, the new gene will need to acquire a promoter; it has to start getting transcribed.

It has to have, you know, it will probably end up getting mutations, and these mutations include insertion, deletion, or whatever. At some point, this should create an open reading frame such that whatever you know is being translated. That amino acid sequence, I mean, leads to something important or some useful product. But that is not an easy process, right? It is pretty clear that lots of things have to come out in the correct way.

And that is why, for the longest time, people thought that, you know, this is not going to work.

But when people started looking for it, they slowly figured out that new genes do arise in this particular way. So, when this happens, typically you are going to get genes that you are not likely to find anywhere. Because obviously, these are not genes that are being transmitted vertically. These are genes that have arisen de novo, and this is what can lead to species-specific genes.

Do we have any examples? We do. So, for example, in Drosophila, there are three genes, you know, CG30395, which is involved in insemination. CG9284, which people have figured out is a non-coding RNA, and CG31882, which has some function. In terms of sperm, you know, sperm, yeah, the birth of sperm, you know, which basically functions in sperm, sorry. So, it turns out that if you knock down any of these three things, then it is lethal. which straight away tells you that whatever function these genes are performing, those functions are essential.

People are still trying to figure out precisely which proteins and how those proteins are working. But it is clear that new genes can be formed in this way. The fourth thing is, of course, horizontal gene transfer, where the gene is coming from something. which is not within the species from some completely different species you know or sometimes an organism of the same species, but not your parents. So, it's not a vertical transmission very common in prokaryotes, responsible to a great extent for all the antibiotic resistance that we are seeing around us. And you know it is a new crisis that is just waiting to explode. I am not going to talk about it. We have already spoken about it. But what I am going to talk about is something slightly different. So we looked at many ways for new genes to come about.

Which one is more important? Which one is least important? What are their relative frequencies? Now, obviously, it is very difficult to give one answer that will be, you know, true for all species. This true answer will most probably differ across species; it will most probably differ across taxonomic groups. However, if we can look at even one case, we will just get some rough estimate, and that is what was done by Zhou et al in 2008. When they looked at, you know, the whole genome sequences of six closely related species in the Drosophila melanogaster subgroup. So, these are like melanogaster, simulans, yakuba, etc. So, these are all very closely related to melanogaster. So, that is why they are called the melanogaster subgroup; but they are different species. So, they looked at all the new genes that are present in these species and tried to figure out what the proportions are and when they did that, this is what they found: it duplicates about 44.1% of new genes. De novo gene origination accounts for about 12% of all new genes, while retroposition accounts for about 10%. And chimeric gene structure through exon shuffling or whatever mechanism accounts for about 30% of new genes. And they figured out that the rate of formation of new functional genes is about 5 to 11 genes per million years which is actually very high. And if you look at copy number polymorphisms, they were able to find about 44.4% of it. So, what I am really trying to tell you is that none of these numbers is insubstantial. It is not like 0.1% or 0.01%, or something like that. Sure, duplication is the major mechanism, but this is 12%, this is 10%, this is 30%; these are all largish numbers.

So, all that this suggests is that barring horizontal gene transfer, All the mechanisms have actually played a major role in the formation of new genes in Drosophila. And that is a very interesting piece of information. One now really needs to do a similar study in other groups and see what the commonality is. Maybe in other groups, again, we will see that all the mechanisms are important, or maybe some mechanisms will be.

More important in some groups, and some other mechanisms will be more important in some other groups. Who knows, right? Now, up until now, we have spoken about how new genes arise. And we have seen that although in some cases the rates can be pretty fast, they are still, you know, in terms of Millions of years, 5 to 11 genes per 10 lakh years is about, you know, 1 per 2 lakh years or so, and that is a long, long time. And even, yeah, I mean that is what I mean, you know, even with that, it is a long time. And although it is okay on an evolutionary scale to gain a new function, this is obviously not going to work when an organism is trying to adapt to ecological or environmental changes, which obviously happen on a much, much faster time scale. So, although it is okay for biological novelty, you know all these mechanisms and organisms when they are trying to, you know living on a day-to-day basis, they need things that are much faster, at least on a generational time scale. So, what are those mechanisms? How do they operate? And you know how does having those mechanisms, How do they connect with all these mechanisms for generating novelty that we saw in terms of long-term evolution? That is the question that we are going to answer in our next discussion about faster timescale mechanisms. See you then. Bye.