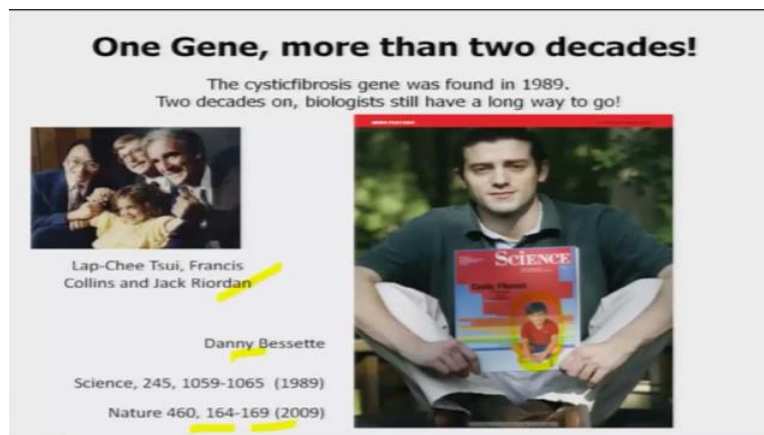


Human Molecular Genetics  
Prof. S. Ganesh  
Department of Biological Scientists and Bio-Engineering  
Indian Institute of Technology, Kanpur

Module - 04  
Lecture - 13  
The HapMap Project

Welcome to the second lecture of the fourth week of this course human molecular genetics. In the previous lecture, we looked into the approach that we normally use for cloning genes involved in recessive disease. So, we have looked at a conventional approach what they call as the positional cloning

(Refer Slide Time: 00:36)



So, what is shown here in this slide is a picture of a person, who is holding the Journal Science. The issue Science appeared in 1989. So, that is the issue that carried the first report of successful positional cloning. So, the gene that they have identified for a condition called as cystic fibrosis and this was found in 1989. So, this is the issue that 1989 issue that carried this story. It was like a, you know, major landmark in human genetics, human molecular genetics, because they were able to identify a gene without really knowing the function of the gene, really based on its position. So, that is a landmark discovery and this approach what we discussed in the previous lecture, positional cloning, has since then, has resulted in the identification of a large number of genes implicated in hundreds and hundreds of monogenic disorders. So, it has been really 25 years, 26 years since the discovery of the cystic fibrosis gene by this approach and what you see is that, you know, now still there is a long way to go.


So, we are able to now clone the gene involved in a monogenic form, but still it is a long way to go. I wish to show this picture, because the person who is holding this issue of Science is the same person, who's photograph is in the issue of that particular Science, because he is Danny who happened to be a patient suffering from cystic fibrosis gene and he was one of the index cases that led to the discovery of this particular gene. So, that is the picture along with the group leader Lap-Chee Tsui and Francis Collins, one of the person involved in human genome project initiative along with Jack. They, you know, really celebrated the discovery of the gene using this approach; it was so powerful. So, this person now has grown up; the handsome individual that you see here is holding that issue and there was a cover page article that appeared in the Science Journal Nature in 2009 to mark the 20 years of this approach since the first gene was identified by positional cloning.

In the two decades a large number of disorders have been dissected, genes have been identified. It is a remarkable achievement in short time, but, you know, still, you know, understanding the functions of the gene that are defective in any disease that you identify using positional cloning, that is still a long way to go. Still, you know, there are several studies, still people are trying to understand the gene and, and this kind of approach, positional cloning really helped in prenatal diagnosis and, and in expecting whether an individual, in predicting whether an individual would have the disease or not and so on. But really has not changed away the therapy, that part is not yet changed dramatically, but really the diagnosis and other things have helped and people are able to save many and so on. So, that is the remarkable discovery. What had happened during this period, since 1989 to now is, is also a revolution in terms of the technologies, in terms of the ways by which you are able to analyse the genome. One of the major changes that happened during this period is the completion of human genome sequence project.

(Refer Slide Time: 04:35)

**Post-Human Genome Sequencing Era**

Human Genome Project



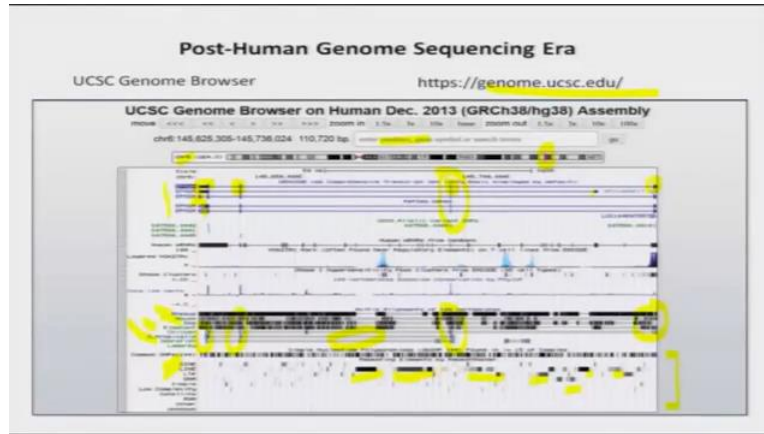
A rough draft of the human genome was completed in June 2000

- 3 billion bases
- 20,000 - 25,000 genes
- 3000 nucleotides/gene (on an average)
- 99.9 % nucleotide similarity to each other
- 99 % nucleotide similarity to chimpanzees
- Less than 2 % codes for proteins
- Chromosome 1 has the most genes
- Chromosome Y has the fewest genes

So, in 2000, we have completed the first rough draft of the human genome. There are two issues; one that came in Science, the other one in Nature, which really detailed the different aspects of our genome. How complexities, how, roughly what could be the number of genes, what are the repeat elements and how the genome, genes are clustered in chromosome, which chromosome has got more number, is there any chromosome specific architecture in terms of genome sequencing? There are many, many aspects that have been discovered using this approach and what we know is that it has got close to 3 billion bases. In the beginning people have been expecting there could be a large number of genes that our genome could have. Now, the current estimate is may be 20 – 25 thousand genes; that is the maximum, still we do not know all the genes and on average per gene it could be 3000 nucleotide and so on

So, what we also know is that it could be 99.9% similarity between two individuals and then we also share about 90% similarity with chimpanzees, our predecessors and then, less then 2% of our genome codes for protein and of course the largest chromosome, chromosome 1 has got large number of most of the genes and chromosome Y is something that has got the least number of genes. So, these are some of those, you know, bullets with regard to the human genome that we know.

(Refer Slide Time: 06:05)



So, as a result, now we know a great deal about our human genome, not only about its sequence, but we also know about what are the genes, where they are located, are there any different splice variants and so on. There are beautiful tools available for you people to go and explore, I am just listing two of them. One is the UCSC genome browser and you can see that the website is here and you go and then you will be able to, understand how the genome is. You simply, you know, in this region, in the query box you add any gene that you discussed in this class or anything that you want to understand and then it would show some display like this. What does it really show? For example, if I give a query called EPM2A a gene that I have been working on for more than 20 years, what it would do is that it would display something like this.

What it really shows is that this gene is located around this region of this chromosome, chromosome 6. It shows exactly the position, where it is and it also shows the cDNA. What is shown here is, you know, this is exon 1, exon 2, exon 3, exon 4 and then you have different transcripts having alternate promoters, alternate exons and so on. So, it gives you the information. Not only that, it also gives you the similarity, sequence wise with the genome of various other animals. For example, rhesus; the genome sequence is compared with rhesus monkey, with mouse, dog, elephant, chicken, xenopus, zebra fish and many other fish, right, lamprey. What you can see here is that there are segments in the genome that show very high similarity. Wherever you have exon, you have very high similarity, wherever there are non-coding sequence, you do not see much of a similarity.

So, over the time during evolution things have changed between the genomes but if you compare with the closest species like the rhesus monkey, you will find that even in the non-coding region where there

are introns, you see higher sequence similarity. That really shows the foot print like, you know there are regions right and it also shows what are the repeats that are present. See, there are repeats called LINEs, SINEs and so on. So, these repeat are present all over the genome and you can see that there are several regions you have the repeat. So, remember we discussed that at times regions of the genes gets deleted and it is recurrent, meaning it happens again and again and it could happen because of what you call as non allelic recombination, because there are repeat elements and during recombination, the two alleles, homologues have to come together and then recombination, at times it can, misalign; as a result you could have non allelic recombination. This is something that we discussed when we were talking about the different types of mutations and you can see that where are the regions that are, having the repeat region and what is known is, for example in Indian population, this particular exon of the gene, gets deleted and you can see that, have repeat elements flanking these exon, which may possibly result in such kind of deletions. So, these are some of the information that are available.

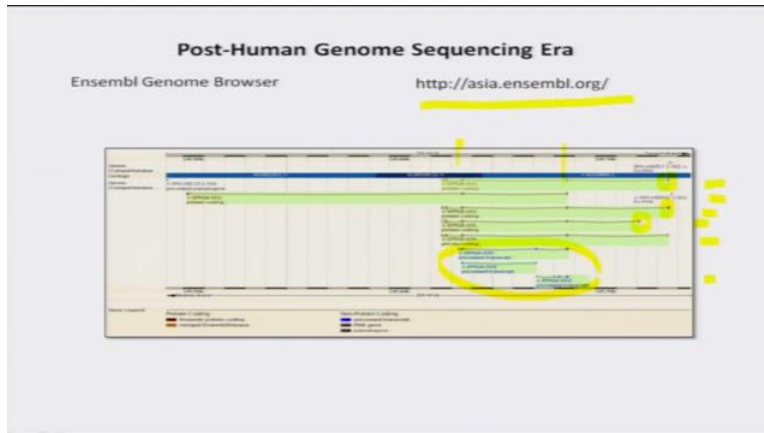
So, if you are a, you know, if you are running a lab or doing a research on a particular gene or its mutation, if you found any deletions you can immediately go to this genome browser and understand how complex that region of the chromosome is. You can look at all these things. So, if you click each one of the heading here, it will give you more information and you can analyse. So, that is a very, very powerful browser which gives you all the information based on bioinformatics analysis and you also see here for example there are mountains that you see and these are regions where they are looking at the histone modification. You know, you talk about chromatin, change in the chromatin, right? You know, there are, for a gene to be active or inactive, you know, you have the DNA along with the histone protein which forms the chromatin, but the chromatin is very compact or very loose. It depends on whether a gene in that region should be active or not active and people even looked at what are the regions you have such kind of modifications in the chromatin, right and such modification can also be modulating the expression of the genes which also you call as epigenetic something, that you know is coming up in the field.

For example, although you would be contributing your genome to the next generation and the expression of these genes also depends on what kind of diet you take. If I am a diabetic person, the way some of the genes are going to be expressed in the next generation is modulated by such kind of modifications we believe. So, that is another aspect and all these informations are available in one click. So, you know, lot of wealth of information available that really helps us to understand our

genomes. So, I would urge you to go and explore and, and try to find. Even, you can compare between different sequences, different species and you can explore what it is.

Another such browser that gives you that kind of information is Ensembl. This is from Europe, the other one is from California, Santacruz.

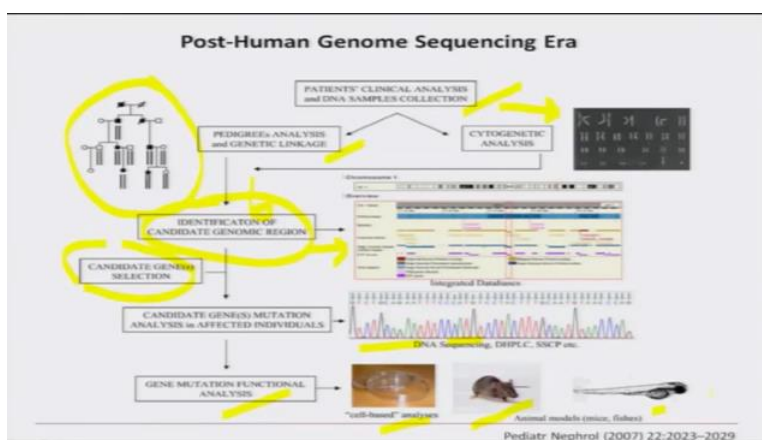
(Refer Slide Time: 12:00)



This is from Europe, which we call “Ensembl” genome browser and again I have given you the site and again it gives you information, very different kind of information, you can go and explore. Lot of, you know, interesting aspects that you have in the browser, a large number of aspects that you want to explore, you want to look into the how genomes are similar between different species, what are the gene rich regions and so on and what I am going is, giving you is an example for a particular gene. Again it is a EPM2A gene. I have given that as a query and it gives you beautifully as to how this particular gene is making various different types of transcripts. So, there are transcripts that has got exon 1, some has got alternate exon 1, exon 2 and 3 again is common and then of course, you have various transcript that show different kind of combinations. So, that is, you know, is something which is very, very important and informative for people who are studying a particular gene. So, now it is all based on bioinformatic analysis and they keep getting data from various publications, try to integrate here and then maintain it. So, it really is something that is for anyone and it is a free site. You go and explore, you will understand lot of features about our genome. There are help files, there are introductory topics and we can download many of the papers and it is really worth going and exploring in these two sites.

So, this is what, now we have discussed positional cloning, wherein, you know, we have discussed how the markers we used to narrow down a region and then you go and look into the genomic fragments, rearrange the marker according to physical order, then identify genes and so on. This approach is pre genomic, before the genome sequence was released. With this kind of powerful browsers, now we really do not need to worry about the physical order of the marker because they are accurately placed in the genome. So, now you do not need to really worry about it. So, if you able to narrow down a particular segment of the gene, region of the chromosome, you can straight away go to this powerful sites and see what are the genes that are present here. So, that is what, you know this, this assembly line of genes discovery, post genomic era is depicted in this slide.

(Refer Slide Time: 14:25)



So, you have to of course start with patients clinical analysis, DNA collections, we have to identify a large number of families and of course pedigree analysis is must; without that you cannot do, we have to look into the pedigree, predict what is the mode of inheritance and of course go and do the linkage studies; take large number of micro satellite markers spanning all over the autosome, if it is autosomal, X chromosome if it is X chromosomal and try to narrow down the region, right? Once you have narrowed down then you can go identify a region, the genomic segment, then you can go and look at the genome browsers just now I described to you. So, you look into region, what are the genes present, it would tell you what are the genes that are there, right and then you can identify a candidate for you to go and sequence or you know, another you know routine approach people do is to do karyotype. Any abnormality even today you start with a karyotype to see if there is any structural change in the chromosome. If there is a change and then, I mean more than one patient who is affected by the same disease, there are identical structural abnormalities likely that it is the same structural change that

results in the disease. So, you look into where the chromosome breakage happened, translocation happened, that gives you a clue, which region. So, that also takes you to this particular step that is to candidate genomic region. Again you go to genome browser, look at all these things.

Then you have to go and identify or select candidate. There may be 20 gene, there may be 30 genes. So, you have to prioritise, pick up some. For example, I am looking at a disease that affects the skin. Therefore, the gene I would consider as a candidate are those genes that are expressed in the skin tissue. If a gene is not expressed in skin tissue, probability that may be contributing to the disease is less likely. So, but it could be a, even a secretory protein, you know, expressed somewhere, protein is exported, it may act on, you know particular tissue. Therefore, when you want to look into, you know, what kind of protein they code for and then consider that as a candidate. So, there are several approaches people use to identify a candidate, line them and then do mutation screening. We go and sequence that particular gene in the patient.

You know, we take one patient for every family and simply you go and screen for the mutations. If you have found some mutation, then you can be confident that is the, mutation or that is the gene that is causing the disease. If you can find same gene being defective in multiple families, all having the same disease, then it becomes much more stronger for you to tell this is the gene likely to cause the disease. But, at the end of the day, you need to validate, the mutation that you find, especially in dominant disorders where the changes are not like what you see in recessive disease; that gene is deleted or nonsense mutation. These are really loss of function mutations. You are certain that causes the disease. But, if you have changes that are changing the amino acid it becomes extremely difficult to tell whether that is the gene which is causing the disease.

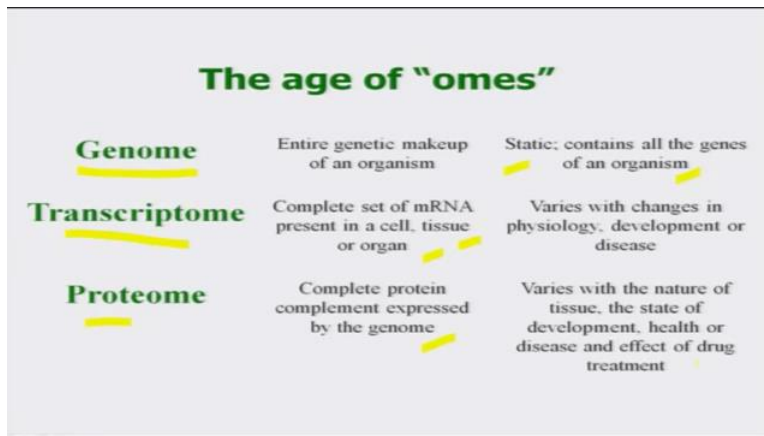
So, you go and do assays. For example, you test the protein or its function at the cell levels, cell biology, create animal models like we discussed in mouse or there are more powerful models that are there; for example, zebra fish, you know, a fish, right, that talks about many of the, you know developmental process. If you are looking at congenital problems, you can easily screen them using zebra fish and so on and if you can recapitulate by mimicking that kind of a change, again in the genome of these organisms and you can recapitulate some of those phenotype defects that you see in humans. Then you become far more certain that these are the gene that causes the disease. So, that is how, you know, these days the genes are being discovered, genes are being identified for disease and genes are being validated for their role in the disease that you are studying. So, this is the flow chart



and now you can understand how things have changed; from simple genomic library you go and do what you call a genomic walking, making contig and then cloning the gene. Now, everything is a click away. Once you have done a mapping in 30 minutes you can identify what are the genes that are present in that region, because of the powerful information that is available to all of us via such kind of browsers that we have discussed. So, you go head and then look into these browsers that will give you more information.

Now, all this information that we talk about, whether these are different splice variants, whether you know, what is the conservation between, I said that there are regions that are, the sequence is very, very similar between say, a segment of the chromosome of fish with human. What does it tell you? It tells you that these are very, very important for the function of the gene. That is why the sequence did not change much. So, this has come, because of a new brand of Science called as genomics.

(Refer Slide Time: 19:52)



In fact this is not only genomics; this is called as the age of "omes", where you have a field called as genome or genomics, which basically study on the genome or the entire genetic makeup of the organism and then you also have what is called as transcriptome. These represent all the transcribed segments of your chromosome and proteome is all the translated product of your body. So, this is what it is. So, what is the difference? The difference is the genome is, pretty much is static. You know if I look into my DNA, derived from my, say skin cell and compare it with, for example my epithelium present in my buckle inside my mouth, the DNA is not going to be very different. It is going to be pretty much identical. But, if I isolate the RNA from by buckle, you know, epithelium, compare it with my skin cell, it is going to be very, very different, right?

So, the DNA is pretty much static. It does not change much with the age or tissue type or anything, except immune cells have very different organization, but other than that you do not really see much of a change. But the RNA which is present even in the same tissue, with age it is going to be different. For example you look at, most of you are youngsters you look at our skin, it will be very, you know, wrinkle free, glowing and so on. Go and look at your grandfather and grandmother. It will be wrinkled, it will not be glowing and one day your skin also would become like that. It happens because of a change that happens in this, in the cell and because of the protein that are made, because of the RNA that you make, because of the genes that express.

So, even in a tissue with age there is a temporal change in the expression pattern, right? It is very, very dynamic. It can change with time. The way the genes are expressed during, for example winter time when it is very cold, skin cell is going to be different from summer time and so on. So, that is about the transcriptome. Then, you have of course proteome. It is, one can say if I have looked at transcriptome, why should I look at proteome? Because, I know this is the gene expressed, therefore that protein should be there. But, it is not simply whether a protein is present or not present. It is also whether the protein is functional or not functional. A protein may be there, but not functional. How do you know that? Because of the changes that happen in the protein, protein gets modified. These are called as post translational modification and these changes are dynamic, more dynamic than your mRNA expression.

So, therefore, you know, there are signalling cascade that happens, there is a cross talk and then protein becomes functional, non-functional and looking at such kind of changes in the protein that field itself is called as proteome. So, people are looking at various aspects, they are looking at the static sequence, the genome sequence, but it gives lot of information because it did not change, now you can compare different population and see who are more related to me or I can look at the transcriptome, the transcriptional status between same tissue, but in different condition and then see what is that, for example infected, not infected, same tissue. I can look into the transcript and see what has changed because of the infection and then decipher whether that change is a cause for the infection or a consequence of it or I can look into the proteome which again would tell you how, what is your metabolic activity and how that may change the way my physiology is and so on. So, that is a proteome.

So, things have changed, people are now looking at global aspect of all the sequence all the protein and try to understand how that varies and how that may contribute to your normal health or you know, in disease condition. So, since we, we are talking about human genetics, so it is most to do with, mostly it is to do with the genome, therefore it is called as genomics.

(Refer Slide Time: 23:58)

**Genomics**

## The age of "omes"

**Genome** Entire genetic makeup of an organism Static: contains all the genes of an organism

**Comparative Genomics**

- Comparisons within a genome
- Comparisons between species
- Comparisons between individuals
- Comparison between cell states
- Integrating genomic information

0016128057197

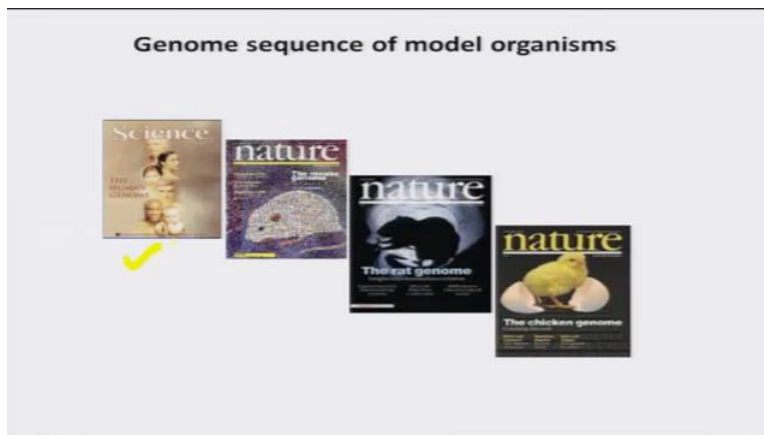
So, people study the entire genetic makeup of organisms and then try to, you know, see if there are changes or variation within individuals, variation between populations and at times between species that would give you some novel insight. This is a new field. So, when you compare a genome sequence with one another, this is called as comparative genomics. For example, comparison within a genome; I can compare, for example chromosome 1 with chromosome 22. Chromosome 1 is very big, chromosome 22 is very small. So, is there any difference between these two? So, why for example chromosome 9 often gets deleted and resulting or translocated, resulting in particular cancer? Is there anything that is unique to that chromosome that makes it more brittle than any other chromosome? So, I can look into the genome of a particular species; within the genome I can understand and extrapolate or I can compare between species.

Now I have the mouse genome, I have human genome. I can compare and decipher why a mouse has got a tail, while I do not. Likewise, I can compare between a monkey and the human and see what difference that makes. Why you are able to study monkey, while monkey is unable to study us? What is that is given in our genome that makes us the most powerful species on the Earth or comparison between individuals? Why I am not as bright as my friend? He understands these lectures better than

me, while I could not. Is there anything that makes him better, because the way the genome is organized, one can look at, or comparison between cell state. A normal cell, a cancerous cell; is there anything that happens in the genome that results in the cancer formation? We know that there are the changes in the genome that results in the cancer and then we can integrate all the information.

We can, for example look into different species. There are species of animal that are known to have cancer, that are not known to have cancer; they never get cancer, but humans we get cancer. Is there anything that is there, signature in our genome that makes us susceptible or at risk of developing cancer, we can study or we can study even how variation in the genome may contribute to cancer or how change in the genome, even within a individual certain tissue, some change happens; as a result you have cancer. So, you can integrate this information and try to understand how the genome may contribute to particular property that you are looking at. That is why it is important that you do not just restrict your sequence analysis only with the human.

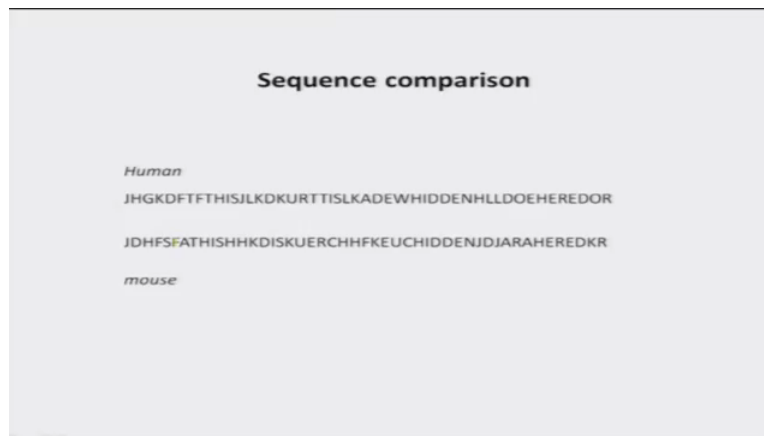
(Refer Slide Time: 26:59)



So, you have, ofcourse the human genome sequence announced that came out in Science, but you have many other species that have been sequenced now. You have mouse, we have rat; people have sequence of the chicken genome is going on. There are now sequence available even for panda the, the species that is so unique to China. So, they have sequenced and so on. So, that gives you an understanding as to how the sequence may contribute to certain inequalities of the species or even to understand how they evolved and this, we can even have some medical relevance. That is where the, sequencing projects have gone on beyond human. They looked at mouse, rat, chicken and so on and go on to understand our genome. The genome is very complex. As I told you, it is only 2% of our genome or less than that may

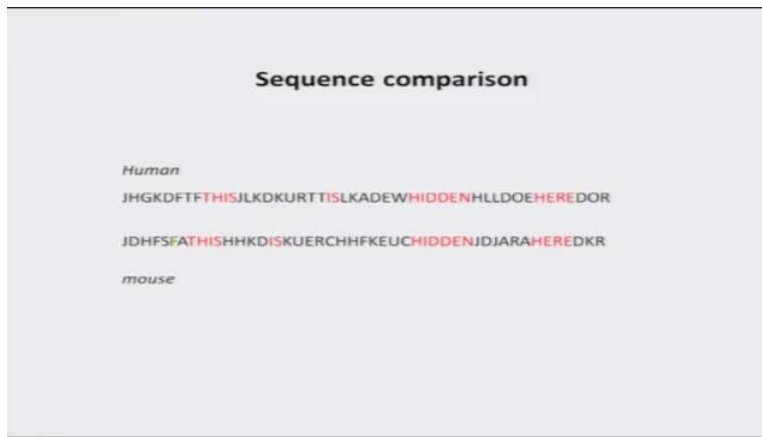
have the gene sequence or regulatory sequence, some functional form. The remaining is so called non coding DNA. We really do not know what is the message that this DNA sequence carry. Certainly they should have some function, otherwise it would have been lost, right? So, they are there, because they have some function, we do not know. So, one of the biggest contributions of the comparative genomics is to identify certain messages that are hidden in the genome sequence, because these are sequence, certain signatures we do not understand unless you compare.

(Refer Slide Time: 28:34)



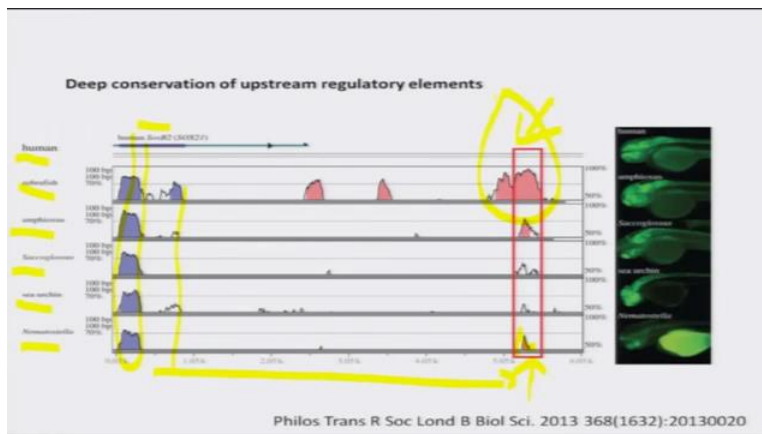
For example, we, I am just, some hypothetical sequence given for, this is a same segment of the chromosome between human and mouse. If you look at it, it would be very difficult for you to understand what is the, is there anything similar between these two, is there anything that is not similar between these two?

(Refer Slide Time: 28:52)



It is extremely difficult, but if you carefully analyse, you could find certain signatures that are identical, something hidden there, right, that really helps. So, this is possible only when you compare two different sequences. If you do not compare, such kind of signatures that are common between two species, you are going to miss out. That is where the comparative genomics really help. These are not coding sequence. These are sequence beyond the coding sequences and they may have a regulatory role in modulating the way the gene, in the genes are expressed or modulating the way the genomes are expressed. So, this comparison between the genome sequences of various animals, model systems with human, gives you some understanding as to how the signatures are possibly conserved or what could be the function of this. So, but you need to validate. That is where model systems really really help.

(Refer Slide Time: 29:48)



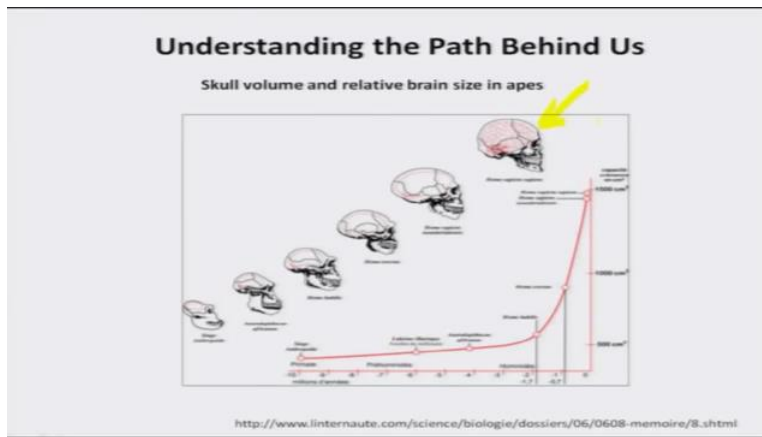
So, what is shown here is such kind of comparisons, is the upstream regulatory sequence of a gene. For example, SOX B, this is a gene. Here the, what they have done is a, you know, a region that is beyond the SOX B2 coding sequence they compared between different species, between human, zebra fish,

amphioxus. This is one of the primitive vertebrates and then many other, like for example echinoderm, and many other lower species. So, what they have done is they have looked at the sequence, they compared and you can see there is a segment which is very highly conserved. The peak here represent that similar sequence and the distance also very critical. This is where your gene sequence is; you can see the coding sequence. This is conserved in all, because coding sequences are expect to come conserved because they have to code for the protein, the codon has to be maintained.

But, what is interesting here is that downstream of that, you still have certain sequence which is not coding, but they are very, very critical. Now you can, what you can do is that you delete this region in the genome and then see whether that development, because this gene is involved in development and it is affected, for example, in any of the developing model. For example, you can use zebra fish and then show whether that changes and even people have tried to swap it, for example use the human DNA sequence over there and see whether it is able to drive that gene and this is what shown here beautifully in this paper. You can go and look into as to how comparative genomics helped us to identify certain controlling element and how the model systems helped us to understand what is the function of them.

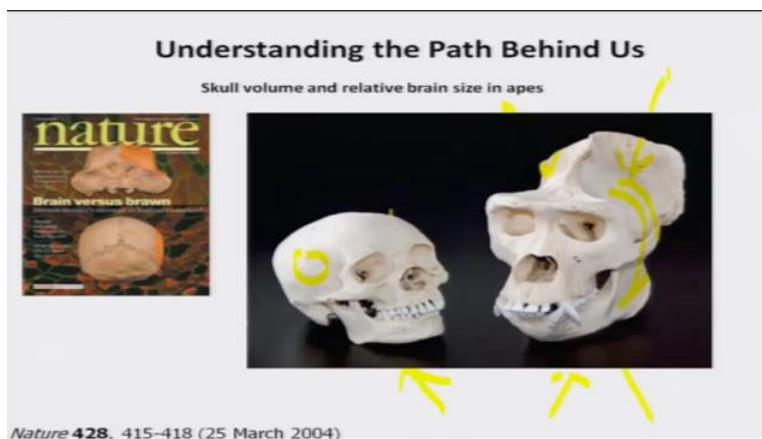
So, now it is very, very obvious for us. If this is, one of the gene is defective in a disease, now what you will do? You will not restrict your mutational screening only till the coding sequence. Now you know this region that is present downstream of the gene is very critical for its function. So, now I would also look for mutations there in. So, if this region is deleted, may be the gene does not express. So, that is how such kind of analysis gives you a power in deciphering how changes in the genome can alter the way the gene function and how they may contribute to the pathology. So, the human genomics is not restricted only to disease that is only when you are not well or having a disease that I am going to bother about you; this goes beyond disease. We also look at the wellness like why you are so successful as *Homo sapiens* as compared to our predecessors, right?

(Refer Slide Time: 32:39)



I am going to talk about one such example which really helped us to understand how we became so powerful species. This is by looking at the genome and comparing it with a property that is skull volume. So, if you look into our closely related primates, you look at gorilla, you look at orangutan, many other species they are really, really mighty. They have a huge head and they are very physically so good and you would assume that they have large brain, because they have a big head. But, that is not the case. If you really compare the skull volume relative to the brain size, we will find that the *Homo sapiens* that is the humans, we have the largest volume as compared to our closely related species. Why?

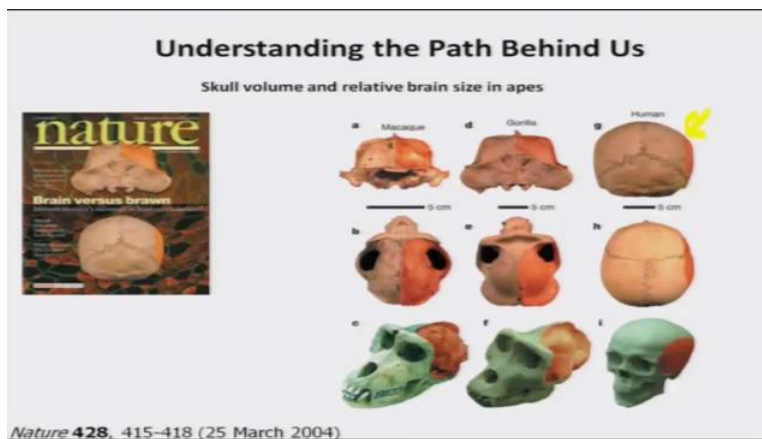
(Refer Slide Time: 33:24)



It is because of this particular property. So, what we have shown here is the skull of two different species. On the left is the *Homo sapiens*, the human and the right you have one from apes, right? The skull is big, but what is, you can note here is that on either side of the skull you have a huge cavity, right? As a result, the inner space available for the brain is compromised in the skull of the apes. Why

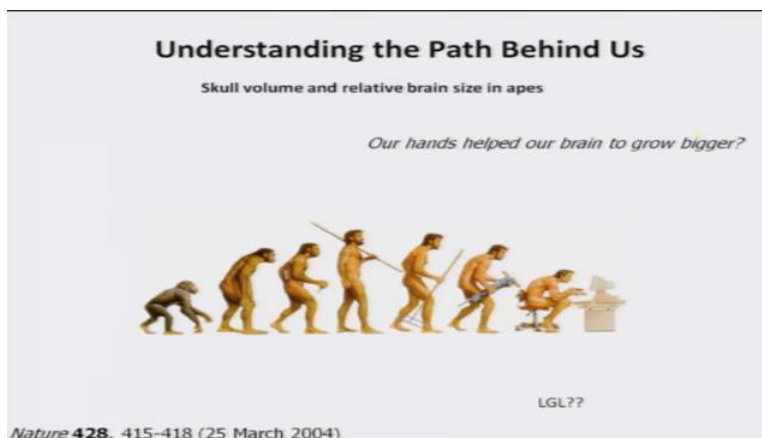






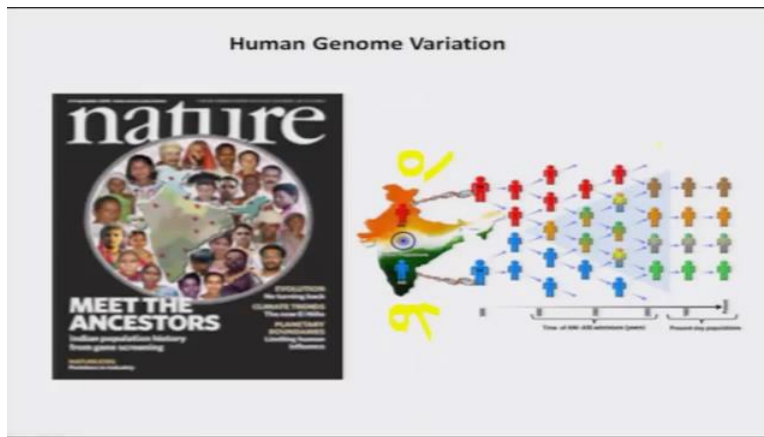
So, this is again an example to show how for example this human brain is able to allow the human skull is able to occupy or allow a larger brain to be present in almost same kind of skull.

(Refer Slide Time: 35:48)



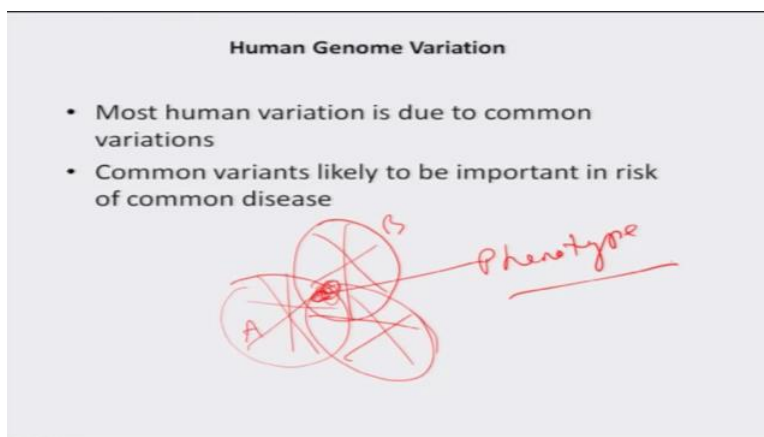
So, this is a kind of an example which tells us that may be we lost the strength of our jaw, but as a compromise our skull now allowed a larger brain to be located. As a result, now we are able to think very differently, we have the ability now to make tools and take care and probably we have evolved the way we are now. So, that is also talks about how we evolved from, may be from our predecessors. But, the genome sequence also helps us to understand how we evolved as a population.

(Refer Slide Time: 36:25)



So, this is again a Nature publication, which talks about the human or the population, Indian population. How different Indians are as compared to the other and they looked at the genome sequence, variation therein. The hypothesis is that we have had two population to begin with, one what is called as ancestral North Indian on the Northern part of India and ancestral South Indian population and all the population that you, say, see today are all derived from the mixture of these two. So, this is likely the scenario based on such kind of powerful genome analysis. So, that is kind of out come of such kind of analysis. So, how do you really come up with such kind of suggestions that the Indian population had such two ancestral population, what you call as ancestral South Indian or ancestral North Indian? This has come from the variations that our genome has. Remember, like we said that, you know, each one of us are very different from our, you know, every other individual, right, very different.

(Refer Slide Time: 37:42)



You have a unique genetic identity and this is because of the common variation that our genome has. We have rich variations in our genome and these variations not only gives us uniqueness, a property, but also can serve as important risk factors for common disease. For example, this is a classroom, 40 of you, like you know, I have 40 students. One of them sneeze; in two days, like I will see another 10 of them get the cold, right? Why others did not? So, it could be because of the variation in your gene, genome that gives you resistance for some and that may give you susceptibility for some. So, that is what we talk about common diseases. So, when you talk about common diseases, these are complex diseases, meaning polygenic. Your disease is because of multiple genes and each gene has got multiple variants. A combination of variant could result in a disease. How that is, you know, we can, we can explain here.

For example, you have gene A and the gene A for example, could have several different alleles and this could be one of the allele that contributes to a disease. Likewise, you have gene B that could also have several allele. This is B and this allele contributes to that. So, when you have more than one gene and one of the allele come together, you may have a phenotype and this phenotype is resulting from the contribution from three genes. But, this is a simplistic model; it could be more than that. In addition that, could be environment. For example, you know, when somebody sneezes a bug (micro organism) that is coming to me; that is environment. So, I may have a combination of the genes that may give me resistance. On the other hand, if I have a combination that gives me risk, so I may or may not develop disease until unless I am exposed to the environment; unless the bug (micro organism) comes, I am not going to express that phenotype. So, that is what we call as a complex disease.

So, how do you really identify what is the risk factor? So, that is the next challenge that we have in human genetics. But, we are in the cross road. We have certain technologies already developed, we have certain approaches already devised, but it is not as successful as what we have seen for the monogenic form, right? This is developing, may be in 5 years, 10 years we are going to see several breakthroughs, with regard to what are the gene combination that results in common disease. So, we will look into some of them. What is the road map? How people are, what is the model and how people are trying to identify this.

(Refer Slide Time: 40:24)

**Human Genome Variation**

- Most human variation is due to common variations
- Common variants likely to be important in risk of common disease

Paradigm shift Human Genetics:  
*Enumerate all common variants*  
*Correlate with disease*

So, the hypothesis is that the common variants likely to be important for common disease. So, if that is the hypothesis or model, what we need to do? We have to identify the common variants. The common variants could vary from population to population. Therefore, you want to identify common variants for all the populations and then, correlate these variations with the disease and see whether these variants contribute to the disease.

(Refer Slide Time: 40:51)

**Human Genome Variation**  
**Mutation Vs Variation**  
Both represent change in sequence of nucleotides

ATAGCTCGCTAG  
ATAGCCCGCTAG

Mutation      Polymorphism

Variation seen in a few individual      Variation that appears at least 1% of the population

So, what is so different, as compared to the Mendelian form? There too, we looked at genes or DNA and sequence variation there and how it is different from the so called complex disease? It is similar in terms of the change in the DNA. So, in either case you have a wild type that is shown here and then you have a mutant or a variant that is shown on the line below. The difference between the common variant and the mutation is, you have mutations and these alleles, the mutant alleles are seen only in a

few individual who are affected or seen only in the affected family. You do not see it in the normal population; it could be extremely low in frequency.

On the other hand, you have variants which you now call as polymorphism, which are present in at least 1% of the population. If there is a class of 100 students and if you are looking at an autosome, you are looking at 200 chromosomes, of which therefore at least 2 individuals or 2 chromosomes should carry that variant. It is very, very high frequent and these 2 individuals may be normal. They may have this variant, but still they are normal and that is why you call it as polymorphism, normal variant and so on. So, that is the difference. It is, at the end of the day it is the difference in the DNA, but how frequent it is, whether it is restricted to few families that are affected or present in the normal population, you can make the distinction whether it is a mutation or it is a polymorphism or common variant.

(Refer Slide Time: 42:49)

**The HapMap Project** 2002

**The International HapMap Project**

- ❑ To determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain.
- ❑ To determine one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe

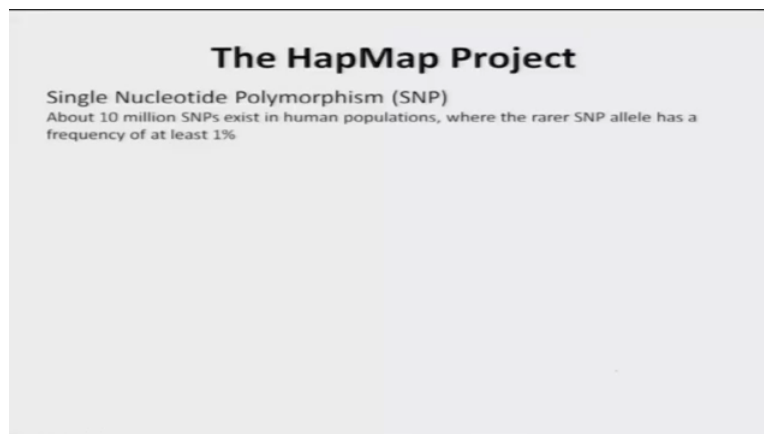
Population chosen: African, Asian, and European ancestry

To identify such variants, the so called common variants which is present at least 2% in a population, you know, a new consortium like project was initiated that is called as HapMap project. Here, the idea is to identify all the common variants across different population. It was initiated it 2002 and this was called as International HapMap project. The scope is or the objectives are to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available. That is one of the clear mandate. It is not something like a company doing this analysis and keeping to themselves. They spent lot of money and made this information available for everybody. So, even a smaller research group or diagnostic labs can use that information to get information and update their research and so on and then the other objective is to determine one million or more sequence

variants, large number of them, their frequencies and the degree of association. Like it is not only you identify variants, but you need to tell in which population you have this allele more frequent and which population you have this allele less frequent; that is very important, the frequency and then to tell their associations. So, you may have identified 50 different such variants, right? So, but how they are associated with each other if they are represent in the same sample and how they are associated or distinct from one population to the other. So, this is what the objective.

To fulfill these objectives what they have done is they have selected populations from all the three continents, Africa, Asia and Europe and they have characterized a large number of individual. Here, if you recall the human genome project, they did not sequence many individuals. Basically six individuals they took and different regions of these individuals they sequenced, but for HapMap project they have done a large case sequencing for especially the coding region of the genes from hundreds of individual representing each one of the population from Africa, Asia and Europe. Again there is, you know, if you give a Google search called HapMap project, you will go to the site where we can look into the variation. For example, any region of the chromosome we can look at, what are the variants known, how different they are in different population that gives, you know, information.

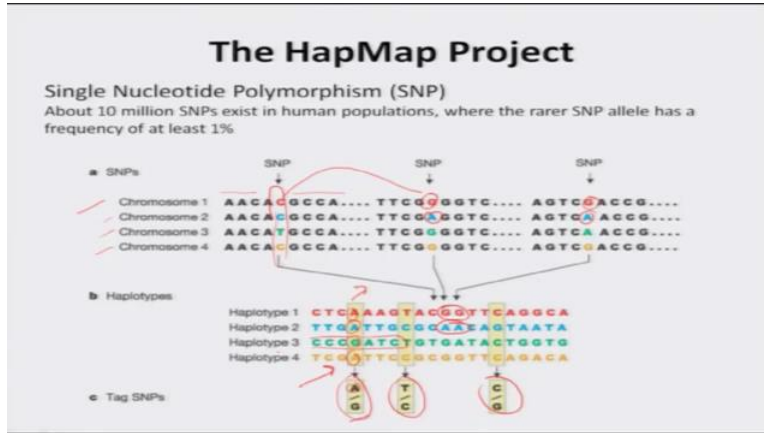
(Refer Slide Time: 45:15)



One of the major focus of the HapMap project is to identify what is called as single nucleotide polymorphism or SNP or SNP, the other way people call it. Why they have done it? Because, if you look into especially the coding sequence, because we assume that most of the risk factors are possibly around the coding sequence, because that may confer some risk for certain disease, therefore the coding variants are more important. Therefore, they looked at single nucleotide polymorphism, because often

you will find a particular base being replaced by other and this new allele is present in the normal population. So, the objective was about 10 million SNPs, you know, we know now that exist in the human population and they tried to understand what are the alleles that are rare and what are the alleles that are frequent at least 1%. So, 1% above is normally considered as risk factor for many of the diseases; so, that they have, they were able to map it.

(Refer Slide Time: 46:18)



Now, you have such information available in the browser if you go to HapMap project and what is interesting here is that this is what, you know, something that tells you about the SNP. So, we are comparing, say 4 chromosomes - chromosome 1, 2, 3 and 4, right? It is sequenced for a particular region of the chromosome and you will find here, you know, these are sequence that are identical, but, but this particular region you have alleles. The allele could be either C or T, right; so, two different alleles that you have. But for another site that are somewhere close to, near by, you have again either G or A. In another site, again near by, it could be again G or A. Now, when you have three such SNPs, SNPs close to each other, now they, you are going to look into a condition wherein a given base, for example C may more often than not, may associate with G or A and so on, right? So, these are linked to each other. So, if I have to do some analysis, am I expected to sequence every SNP to find what is the variation that is there in my genome? The answer is no.

The HapMap project, what they have tried to do is they try to look into a large number of such SNP and then they are able to even tell, for example what kind of combination they have. G probably more often associate with G, A probably more associate with A and so on and then, they came up with some very informative, you know, SNPs like what is called as Tag SNPs, which if I type, you know, I type this



site, this site and this site, these three SNPs would tell me what are the other flanking sequence likely that I am to carry. For example, if I have A, so I am going to have this, these two, right and if I am going to have G, then I am going to have this combination. So, this is something that is called as haplotype, that we described already.

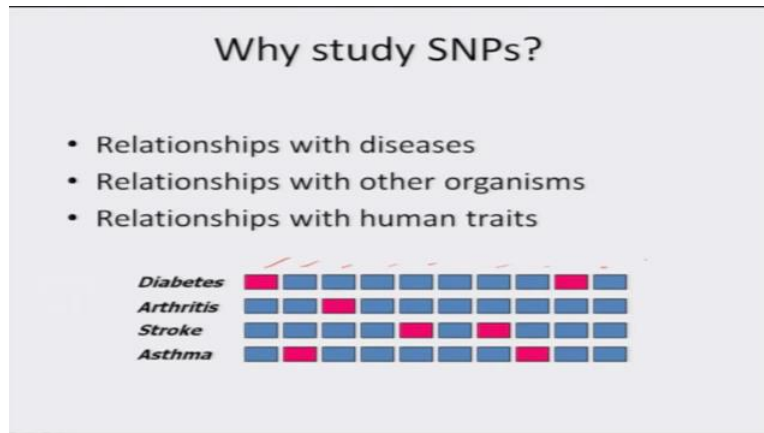
When you talk about monogenic form, there are, you know, physical linkages, there are some alleles, they are linked to each other and they segregate as a block. That is again is a kind of indication here, they are segregating together. So, to identify what SNP I could have, I need not type everything. I can type one of them in a block, which would sort of tell me what are the SNP I am likely to have. So, that is one of the major advancements in the HapMap project. They identified what is called as the Tag SNPs, which helped me to identify the region, not necessarily every variant that I have to look at it. So, why should I really do this kind of exercise; take hundreds of individuals, sequence them, what is common, what is rare, what is the haplotype, what are Tag SNPs, what is the use?

(Refer Slide Time: 49:17)

Common Variants – Common Disease hypothesis	
Common variants important in risk of common disease	
<i>ApoE</i>	<i>Alzheimer's disease</i>
<i>Factor V</i>	<i>Venous thrombosis</i>
<i>MTHFR</i>	<i>Cardiovascular disease</i>
<i>CCR5</i>	<i>HIV resistance</i>
<i>HLA-DQ</i>	<i>Type I diabetes</i>

The hypothesis is that common variants contribute to common disease. So, if I know the variants, now I can look into individual that have the disease and then ask the question what are the variant that are more often present in these individuals? If certain variants are more often present in these individuals as compared to the controls, they are likely to be the risk factors. That would give me some predictive value, right, so that that can be used for treatment or even diagnosis and many other. There are many genes that, I mean, you know, variations therein already linked to, as a risk factor for disease that are listed here, even before the HapMap project came. But, this HapMap project is, you know, would certainly help us to take this advancement further. Why do you study SNP?

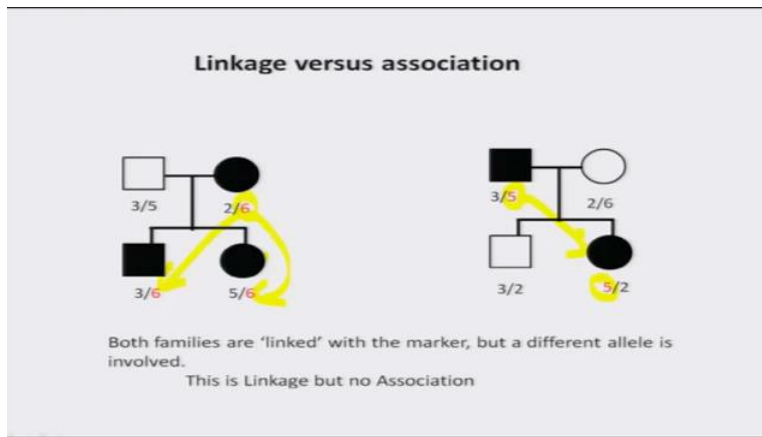
(Refer Slide Time: 50:07)



You want to look at the relationship between the disease, a given SNP or combination of the, you know, more often associated with the disease. You even want to, look at the relationship with other organisms that might help us why we have certain phenotype that is not seen in other relationship with human traits. You know you can even, for example even behaviour can be considered as a trait; then we can try to link. So, this need not be something a disease. Now, being, somebody is being more aggressive, somebody is being more social, somebody is being not that social, these are all behaviour may be something to do with your DNA. But, they are not grouped as something abnormal, but still these are traits. So, one can, you know, correlate them and see whether they contribute. So, what is shown here is that each block probably represent different segments of your genome having different alleles, because of SNPs. Some of them may be involved in diabetes, some of them involved in arthritis, some of them involved in stroke, some of them in asthma. So, when you do this analysis, you will be able to identify certain combination that gives you more risk of developing a particular disease. So that is the ultimate goal.

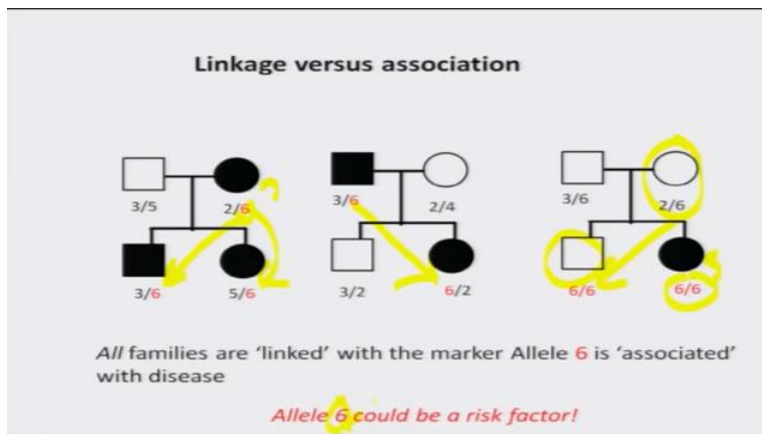
So, how do you do? How different it is as compared to what we discussed in the previous lecture that is a monogenic form? In monogenic form what we have done is, we have compared macro satellite marker, a linkage, meaning that is present in close proximity to a gene that is defective. So, a marker is segregating in a family always with an individual who is affected. We would assume that marker is present very close to a defective gene, as shown here, right?

(Refer Slide Time: 51:50)



So, what we have shown here is two different families. You have a marker and an allele, for example repeat having an allele, which is 6. This allele present close to a gene that is defective. Therefore, wherever you have the 6 going that defective gene also goes in the next generation and you have the individual who is affected in all three generations. Now, the same marker may be linked again with the same disease gene in another family, but not necessarily the same allele. We can see here, this is a same marker, but allele that you see here is 5, a different allele and that allele co-segregates with the phenotype for sure, but it is not the same allele that you have seen in the other family. So, this is the difference. Here, the marker is physically linked with the disease; that is for the monogenic form. But in a polygenic form where your alleles contribute to the disease phenotype, it is very different.

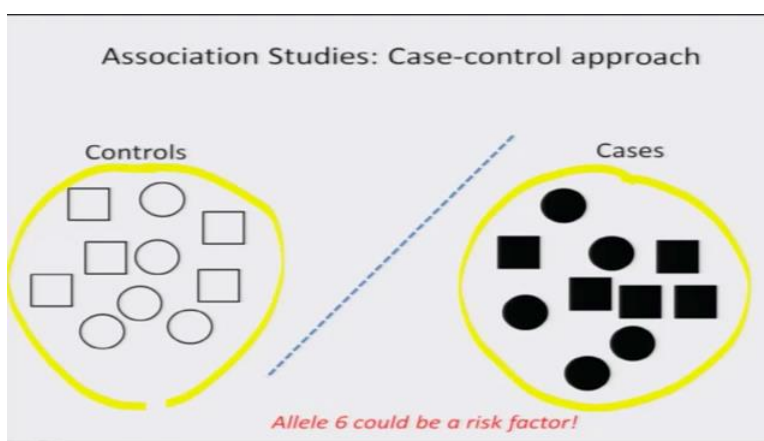
(Refer Slide Time: 52:54)



You can see here there are three different families. Let us say the risk factor is allele 6. So, that gives the risk. You can see here, this individual is affected and this risk factor is contributed to both the, you know, son and daughter, the children and both of them are defective, affected; likewise here, likewise

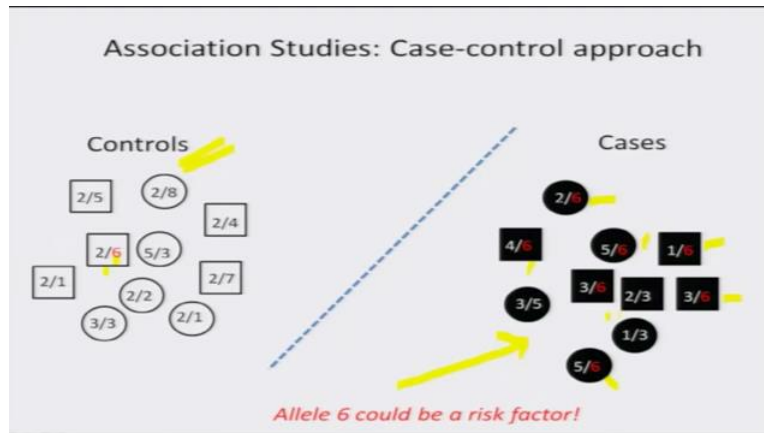
here and so on, right? So, you find that this allele 6, associating with the disease; the same allele. It is not the, you are not talking about the same marker, but particular allele associating with the disease. At times you can have an individual who is having the disease, allele, but normal like here and here. That is possible in polygenic form, because we have, our model is there are many genes that contribute, many, alleles of many genes contribute. These individuals, they may not have inherited the other alleles, right? So, therefore here we are talking about a particular allele that is contributing to the disease. That is the distinction, major distinction between a linkage of marker that we discussed in monogenic form with the risk allele that we discussed about in, in polygenic form.

(Refer Slide Time: 54:03)



So, how do you go about doing, looking at complex disease, because we are not looking at one particular marker or a gene in a family and developing complex disease? These are all many of them. So, the approach is, you group individuals. For example, what is shown here is control, the other one is case. So, all the individuals that have a given phenotype, for example I am looking at a disease and the disease is likely that it contributed by, the disease is contributed by genes, but multiple genes. So, what I do? I group individuals having the disease in one basket, then I group individuals who are from the same population, who are of the same age group, who are exposed to similar environmental condition. So, what I am doing is, I am removing all other variables. I am taking the same ethnic population, because more or less the genome sequence would be similar and then I am taking same age group, because that disease could be age dependent. So, I am taking age, similar age group. I am taking similar number of males and females. So, I pretty much, other compounding factors I have sort of, you know, kept it common between these two. Now I ask, what is the difference between these two groups for the difference in the signatures of certain, you know genes with reference to their sequence variation.

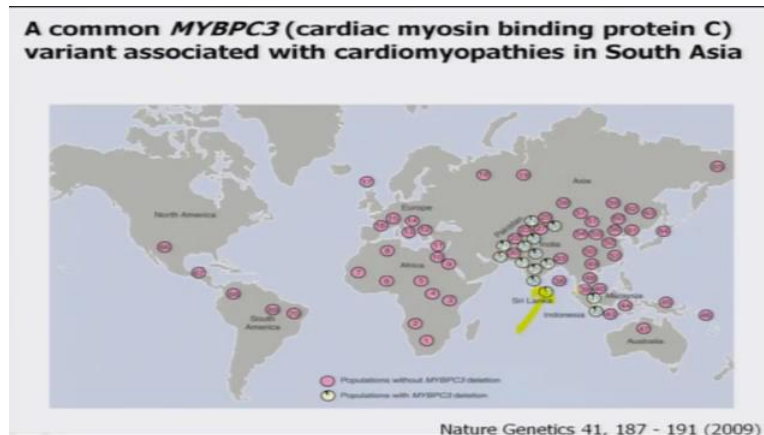
(Refer Slide Time: 55:49)



Say suppose I am able to identify a risk factor, genetic risk factor, a gene contributing to the risk and look for the variants, the alleles, the risk factor, something like that I am expecting. What I am expecting here is, assuming allele 6 is the risk factor, you would expect the frequency of this allele present in the so called controls, who are the healthy individuals, would be very low. As you can see, there is only one individual who has got allele 6, because this is the risk factor. The moment you have, we are at higher risk of developing a disease. Therefore the probability that I would have a disease later is higher, therefore I will end up. That is why individuals that are having allele 6 are more common in the cases, right? We can see that majority of the individuals have allele 6. Still you have some individual who do not have allele 6. It is likely, because for them the contribution comes from some other genetic risk factor, some other gene, some other allele, again you may have the same gene, same disease, right? So, this is the approach people use; case, control and do what is called association.

You see, the increased presence of certain risk factor in the group, the affected, so that is called as association studies. That is the approach currently being used to identify genetic risk factors. This is the first step, like what we did in monogenic form. You map a region of the chromosome and identify a gene there. Whether the gene contributes to a phenotype is the next step. Likewise, here you have an allele that is linked or associated with the disease; whether this is the allele that contributes directly to the disease pathology one has to do a lot of functional studies; that is still a long way to go. But, this is the current practice to identify the risk factor. I will show you one example, which is again probably I have shown you before. This is a common variant, again a small deletion of a small segment of, you know, DNA, which confer risk for developing, you know, cardiac arrest, right?

(Refer Slide Time: 57:46)



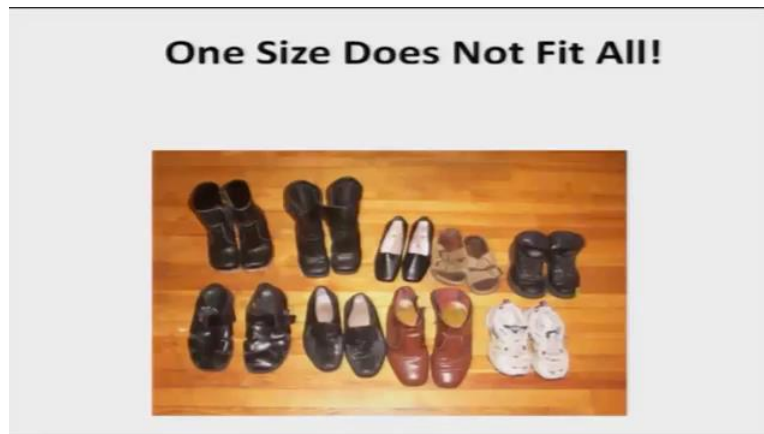
This is something that we, you know, we spoke about in the earlier class also and you can see that that is unique to our South Asian population. You can see that all these circles that are with yellow colour are the populations, wherein you have the particular variant allele, risk allele, like allele 6. So, there is the deletion in one of the exons, sorry, one of the introns that affects the way the gene is spliced. Now, you also have this, this in normal individual. It is not that it is not present. It is present in 5% of the population, in most of the Indian population. 5% of the population has got the risk allele, but not all of us develop. Probably there are other contributory factors, but the, the probability that I would develop the disease, you know, if I have the risk allele, it goes up, you know, two fold, three fold, four fold, you know, increased risk of developing the disease and it could be population specific.

Therefore, one need to really understand the sequence variation in various population. This is very, very important, because there could be, you know, changes that happened after the population has evolved in a given continent. So, you may be having something unique to your population. You can see here, the HapMap project may not have, identified this particular allele, because their analysis does not include, the Indian population. So, we need to come up with our own, sequence analysis. It is very important.

Another aspect as we discussed, disease and trait and so on, is also something related to treatment. That is you know now that in your own family one of your relatives could have had certain disorder, disease, infection, whatever. You go to your doctor and the doctor prescribes certain medicine and he gets better. You know, another distant relative also goes to the same doctor, he gives the same medicine, it does not

really help her and then after one week he goes back, she goes back to the doctor and says, doctor it did not help. What the doctor does? You know, he, normally he will change the prescription, he will give some other drug. The reason being not all drugs help everybody. The way we respond to the drug also depends on your genetic makeup, because the drugs after all work in combination with your biological system. So, drug may go and bind to certain carrier protein, drug may go and activate a receptor, drug may go and change the way an enzyme functions. It is going to interact with your own biomolecules. The, how the biomolecules function depends on what kind of variations you have. The enzyme may be more active, because of a change in an amino acid, your receptor may be responding to a ligand differently as compared to another individual, because there is a variation in the coding sequence of the receptor. Therefore, the way you respond to a drug also is individual specific.

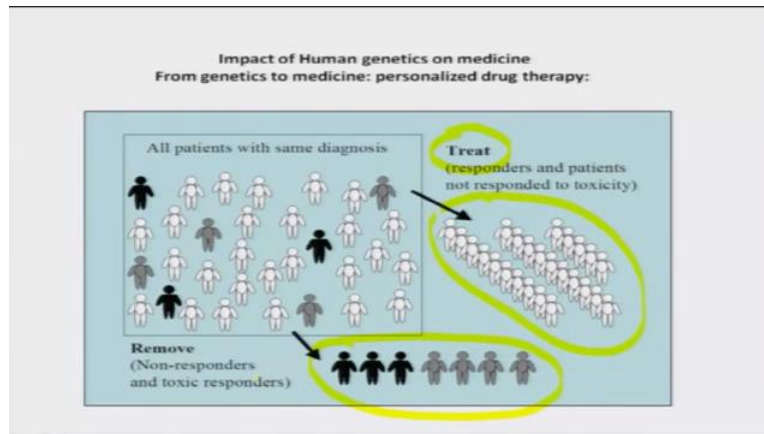
(Refer Slide Time: 1:00:57)



Example is something like the shoes. If you want slippers, you want to go and buy in a shop, you know, what do you do? You know, the shoe keeper, rather the, the shop owner keeps a large number of varieties, different sizes and you try out something that fits you and then you pay and then come back. It is not that every pair fits your foot. The same thing happens to the drug. Not every drug is as effective as it is in another individual, because your genome is unique. So, one of the predicted outcome of the genome variation project is to identify the variants that contribute to differential drug response. Whether, if I am a carrier of a particular variant, then would I respond to the drug the same way as the other person or would I not respond or the drug may be more toxic, if I have the drug it has some allergic function. So, that is probably because of the change in the DNA sequence. So, if I know that I am not to be, I am not the one who would respond to a drug, why I should take the drug?



(Refer Slide Time: 1:02:09)



So, if I can pre screen individuals as, responders who are likely to respond to a drug or non- responders who will not respond to a drug, then you give the drug only to this particular group who are likely to respond to the drug. The other individuals, who know because of the genetic makeup, may not respond you try out, to begin with some other drug. So, this is called as a personalized medicine. That is, the medicine is customized for your genetic makeup, right? So, that is the individual centric; so, that may be the future. So, when you are born, they will isolate little bit of your DNA, sequence completely, profile, they will tell, ok, this guy is likely to develop diabetes, hypertension, and likely to, to have a toxic effect for a given drug or may not respond to a drug. This data is available as you grow up, may know, you are told you are likely to have diabetes. So, what do you do? You better take care of your weight, better take care of your diet, therefore you can minimize the risk of having developing diabetes or you know, you have this data; your doctor looks at the data and tell, ok, this drug should not be given to him, because he may not respond to that or this may be allergic to him, should not be given. So, they go for more informed way, what should be given, what should not be given. So, this is the future.

We do not know all the signatures that give you the phenotype. That is for every drug who would respond or who will not respond, so we do not know. This is an evolving branch of science and this is something expected. So, may be in your own life time we are going to see such kind of revolution, which pretty much, you know, you will carry your chip with you which has got all your sequence information. The doctor is going to put it in the laptop and going to say ok, you guy have to be careful. So, this is something that is expected and this ofcourse sort of primes you to become one of the person who may contribute to such advancements and with that little note, we are ending this lecture, the final lecture of this course, human molecular genetics. We hope you have enjoyed and understood some of



the recent advancement in this field of science and if you have any query, doubts, you may write to us and also do join in the hangouts, we will be able to discuss. Thanks a lot.