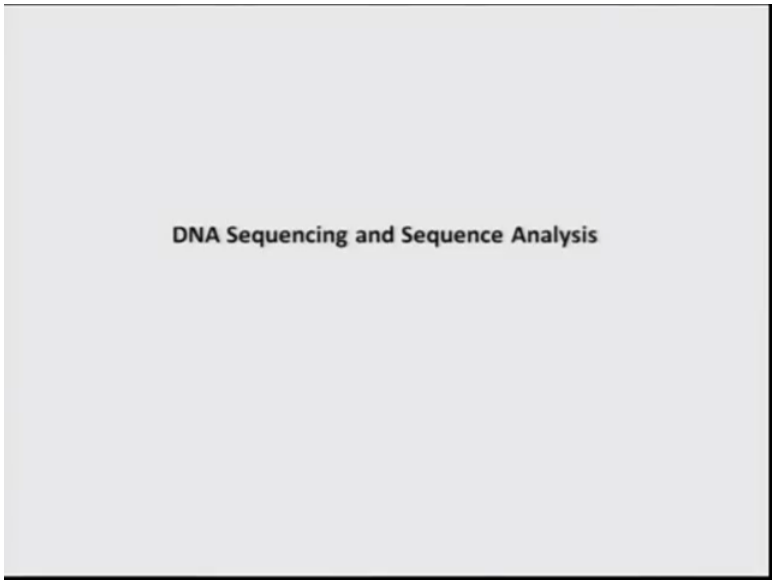
Functional Genomics
Professor S Ganesh
Department of Biological Sciences & Bioengineering
Indian Institute of Technology Kanpur
Lecture No 09
Genome Sequence Database

Welcome to this course functional genomics. So in the previous lectures we looked into the use of microarrays and how they can be used to understand gene expression profile that's high through put approach to look at genome wide expression pattern that was a powerful technique people have developed to understand large scale difference in the expression patterns. Even for those genes that have not been characterised before.

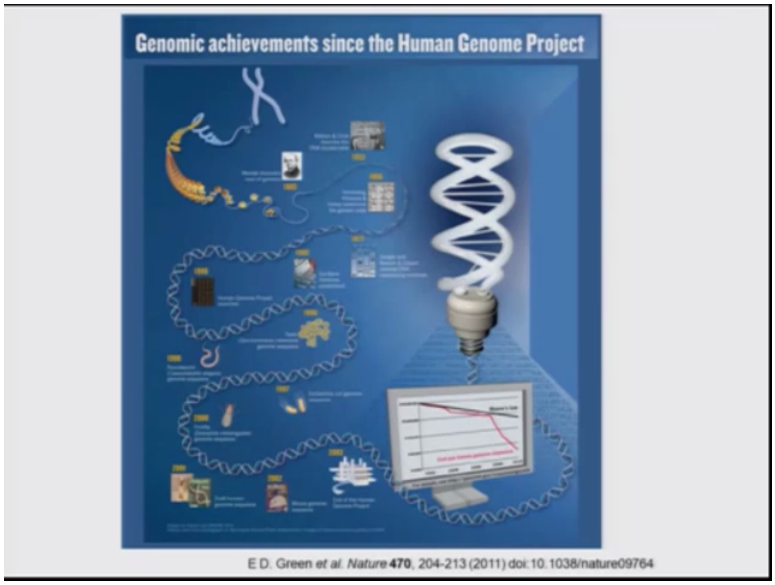
Similarly even understanding the genome the way we look at the genome and that been characterised as change for a time and in today's lecture we are going to look into how the genome has been characterised and the approach is that we use to sequence the genome to understand the complexity that comes along with the sequence and how that information can be used for example in the medicine or understanding how humans evolved understanding how other species have evolved. This has become the challenge and this has been met already with powerful tool that that have been developed.

(Refer Slide Time: 1:28)



So we will be looking into some of those you know advancements in this field.

(Refer Slide Time: 1:33)



So let us look into the slide this slide sort of summarizes how the genomic advancements since the human genome projects that was launched sometime in early 90s have changed the way the genome being studied and understood. So what has been shown is the summary of a topic we already discussed this includes the how the classic example of how Mendel you know sort of understood the possibility of DNA being or the genes being there or how they are segregating and then the early 1919s and up to 1954 or so when the DNA structure was understood and the tools for DNA sequencing have evolved and since 1990s there has been an explosion with regard to how the genome is looked into.

So you can see that in 1990 there was an agreement as to there has to a consortium involving several labs to sequence the human genome with that initiative the methods by which the genome sequenced as a tremendously changed and many new novel techniques have come in and and different research groups and countries have started sequencing several other genomes of moral system.

For example east genome sequence was completed in 1996 and then you have the worm (()) (02:55) genome has completed in 98, in 2000 drosophila sequence has been completed and equalize sequence also complete in 1997 and so on. So with however with the completion of the human genome that has been explosion in terms of the way the genome has been analysed and that is being sort of you know of discussed in the next slide.

(Refer Slide Time: 3:25)

So this sort of summarizes as what had happened since the completion of human genome sequencing in 2004, 2005 it pretty much we have done all the sequence analysis as well and then a large number of sequence have been sequence. For example chicken you know another example very good moral system to understand the biology specially the development biology.

Rat is another very good model for specially drug discovery and dog genome sequencing because this is one that is domesticated and a large number of mutants that you see different breeds are available and that has become again powerful tool to understand what genetic change causes a particular (())(04:03) again that has been sequenced and then you have the closest to the Humans the chimpanzee even that genome is sequenced and with such advancement what has happened is that people started looking beyond just looking at the genome sequencing.

So what you see here in this particular section is the first genome wide association study a topic you discussed a little later but just to give an introduction this is to understand how variation between different individuals can confer a given individual resistant to a particular disease or more susceptible meaning at higher risk of developing a disease and that is what lead to what is called the (())(04:45) project. In this what is being looked at is not the human genome sequence but the variation in the human genome sequence.

So in this kind of developments you know led to even far reaching consequence. For example there are you know when when the human genome sequence is sequencing project was initiated it was a very very expensive offer so governments were involved public money has been put into because this cannot be really done by individual labs or individuals but things have changed with technology that is emerging even first you know direct to consumer vole genome test meaning one could sequence the anti of genome and understand as to what are the variants.

What are te variants exists than what risky could have and a so on so this has emerged in early 2000 in mid-2000 and then different other models for example honey bee sequence is started and then you have with all these advancement sequence and that co-relating with the geno type you also need to have very good data base which can integrate all this data. So you have these NCBA data base for geno type and phenotype that has come in and then other another monkey sequence has come in and ad what you called as again looking into the variation in the human genome.

The world compressed what is called as case control study consortium you know approach is started like you recruit thousands of individuals that are normal from a given geographical ethnic region and then look into individual who have a particular disease that representing the same population and then compare the genome o see what are the variants that make an individual at risk of developing a disease or you know resistance like you know your risk of developing diseases rather low.

So that is break through something which started and then you know the given different for example different countries have started sequencing their own population. Reason being the human genome sequence has come from certain few individual that probably not represent all the population that live on the earth. The Chinese government you now initiated a sequencing of their you know population 2007 and then (())(07:10) and people started understanding how possible the mammals evolved for example so there is a one connecting link platypus that genome was sequenced and there like ways in African population.

The (())(07:24) genome was sequenced to understand how humans evolved overtime and then it went to further investigation especially the cancer genome because we know now cancer is not caused by you know in single gene it is caused by defecting multiple genes and that could be

restricted to the cells that have become cancer and it could vary from one individual to other or when within an individual it could vary from one population to other.

So if you can understand what has gone what kind of changes have happened in the sub population of the cancerous cells in terms of genome than it becomes extremely you know critical in or important in useful in treating the cancer because one it gives an understanding as to what is the subtype or we looking at different groups of cancer or it is this one particular type and second by looking into genes that have been altered you would be able to tell her the duck such you are able to control their individuals.

Such kind of initiative you can see there are many especially if our cancer all (())(08:30) so on. In 2009 again another Asian country the Korean how developed you know programme to sequence their own genome and then it is just not the sequence but beyond sequence how genome has been altered. What you called as epigenetics something that we discussed sometime back, which talks about how.

For example methylation at the chromatin level or at the DNA level can change the way the genome functions. So that is the initiative in 2009 where there was a method by which we are able to identify bases that have been methylated even at the chromatic level you know (())(09:13) modification are been looked at and then some of the commercial animals like species for example bromine genome has been sequenced and then you know you have a large number of mammalian species you know the genome sequence is completed in around 2009.

So we have sort of understanding as to how different the different mammals and how we evolved and so on. So 2010 onwards you know things have changed more dramatically now we have large number of what is called as the association studies have began like people looked into the old genome to understand what kind of changes that or combination of changes can make an individual more susceptible for a given disease and this is a landmark because 500 genome wide association study is published by 2010 that also tells how the approach is evolved over time and how useful it is become in identifying the risk factors and of course you have other countries joining in.

For example southern African genome sequence initiated and there are many large number of such functional genomics approaches also has been initiated one is for example (())(10:24)

limitations you go on imitating about 1000 individual genes in the mouse and understand what phenotype they developed. This something that we discussed in the previous lecture how to understand the function that is from genotype to phenotype consortium and then encode projects to understand the expressions and their correlation with the function and then also understanding how humans have evolved.

Some of the ancestral possible species that once lived connecting the primates with the humans understanding our history evolution and then another landmark project that was initiated was the thousand genome pilot project that was sort of completed, gain here to understand not just the genome sequence but to understand what are the variation that exists between different population here the idea used to complete the genome sequence of minimum 10,000 individuals in a larger context but to begin with it was a pilot study 1,000 genome which was completed which will tell you the sequence variants within the coding sequence non-coding sequence regulatory sequence and so on and these are not restricted to certain population.

They have we will discuss that they have you know they selected population representing the entire human race. So this is you know advancement so we are going to look into what really made a difference therefore we could take up such challenging you know genome projects but such advancements are not restricted only to the other country.

(Refer Slide Time: 12:10)

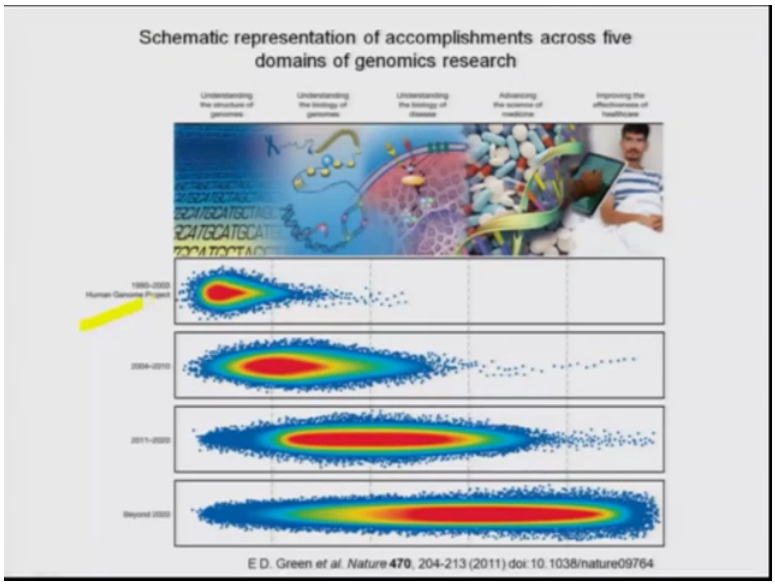
Even within our country India the government of India has funded a project what is called as a Indian Genome Variation Consortium because the Indian population cannot be called as a single population. There are huge variation within different population that are present given ethnic group or the people who speak a particular language and have community, casts , tribes and so on.

And therefore there is a larger gain if you can understand what is the sequence variations that are unique to each one of the sub-population that represent so called Indian population. So there was a consortium mode in large closed to 100 individuals genome sequence o at least at the exome level meaning coding sequence they have been sequenced and that was initiated and you like wise there was a pan Asian population initiative meaning all the Asian countries including India, China, Singapore you know Japan, Indonesia, Malaysia.

They joined together to sequence the and understand the variation within the genome of the Asian populations because there are lot of migration that happened between the population therefore that would one help us to understand the similarity and difference amongst the population and to to understand for example what makes a genome unique and that reflex in terms of you being in a better condition or you being not in a better condition. For example we know that Indians have very high risk of developing diabetes but the risk of developing other disorders like Dementia, Alzheimer are much lower as compared to the western population.

So it looks like that could be something that is there in the genome that make you more susceptible for a given disease or resistance to that. So this kind of approach is would help us to identify the signatures if not the real mechanism but even if you understand the signatures that identifies such sub groups. One could start working on understanding how the signatures really may modulate the cell tissue and organism such that you know you more susceptible or resistance to given disease.

(Refer Slide Time: 14:39)



So this is schematic representation to tell about how the genomic research has changed over years. This is something that that we already discussed the human genome project 1990 and 2000-2003-2004 the draft sequence was completed. So you can see the understanding the structure of the genome in with reference to the human genome at least is done by pretty much is completed now because you know we have sequenced the genome, the sequence is available but what is important is understanding the biology of the genome.

What is it mean so you have a sequence so you have sequenced it? What really it means that is becoming extremely challenging because there was an expectation that when the genome sequence was completed that you would understood the biological bases of you know the majority of disorders you will be able to treat them and so on but it really did not help in that way. We only understood what is the sequence but it has really thrown a more challenges than what has been anticipated.

For example one of the expectations was we will have more number of genes as compared to the other species because we are more complex and more successful but with you know the completion of human genome. We found that we are not really having a large number of genome genes as compared to other species. The complexity is (()(16:00) so to understand that you know the biology of the genomes you know how do they function. It has become challenging and that is what we can see here in from 2004 to 2010 such kind of approaches have come in, so we want to understand the genome that is what you call as functional genomics.

So how genome regulates its function right so that is the functional genomics and then from 2000 that is still you know sort of ongoing. You can see that its sort of picked up and we have you are talking about how a genome is organizes dynamics and its transcription variations all these things are coming up but that you know sort of trying to understand what possibly you know contributes to normal function of the genome. The more challenging would be to understand how variation in the genome or change in the genome can contribute to the disease progression genesis and so on.

That is going to be challenging and that is you know is already in progress. You could see that we will talk about it little later. There are number of studies we are understood what is a genetic cause and how that can possibly alter the way genome functions in terms of expression in terms of miss expression and so on. So that is this study currently been investigated and and that would you know lead to much more impact in the way we manage disease that would result in what is called as advancing the science of medicine.

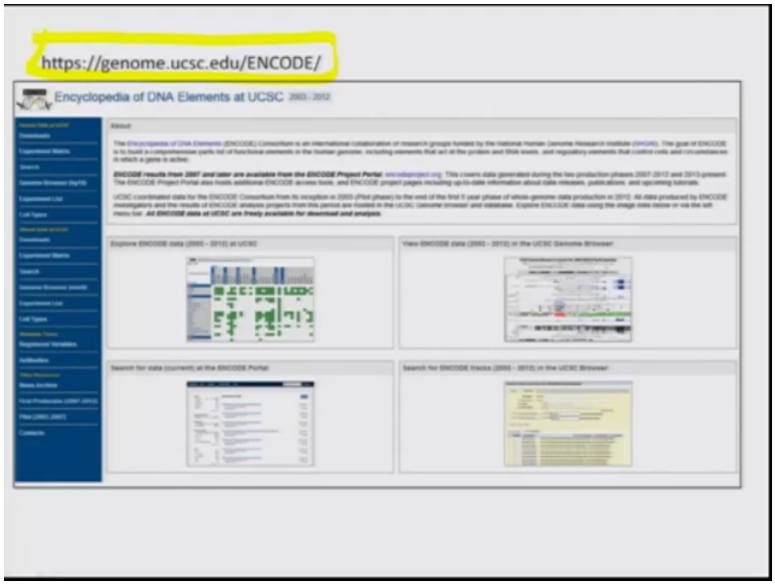
So we right now the focus has been to understand what goes wrong as a result you would have a disease but can we fix that. Can we alter can there be a therapy based on your genomic understanding and that is something that is going to come that is going to be the challenge the next decade and obviously in a such advancement if and when made we are going to have you know improve your health condition.

So it is not that you would change your what you called as life expectancy you may live you know in 90 years 100 years but the focus has been that can we have a healthy living without much of a problem. So this is not the longevity but you know whether your it is the healthy living that is what it is, as long as you live you are happy and you do not have any discomforts that itself is a huge relief to the ageing population because which time with industrialization and increase in the life expectancy.

Over the time we are going to have the aged population the ratio in most of the countries is going to increase including India and 50 years from now is going to be a huge burden because you are going to have a more of aged population. So that becomes a huge challenge for countries to handle, so if you could cut down the expenditure on the medicine than you are able to live better

right because for every-thing is connected to money. So these are the expectations that is how it is projected and we will see some of these issues.

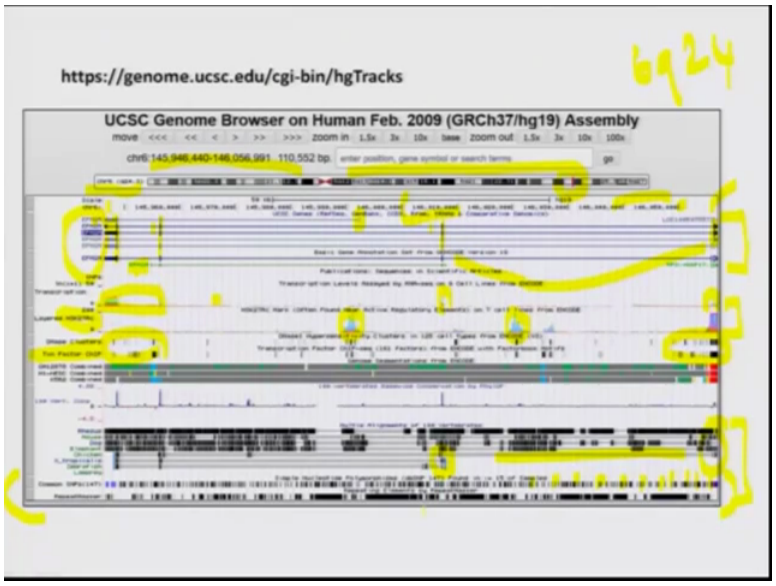
(Refer Slide Time: 19:20)



I am just showing you one such website which I would encourage all of you to go and look into. This is called as a genome browser from the University of California website and that talks about the encyclopaedia of DNA elements it talks about how the sequence of the DNA and it is association with transcription, modification, involvement in disease and everything. So if you really want to understand what the genome revolution has done to our understanding of the genome.

This is one of the very good websites and very educational you can go and there are many good links where you can understand how it has really you know change the way we understand the genome. So one is to of course the data if you are interested in analysing the data but you can also look into some of the applications like for example the genome browser. It tells you how the genes are organized right so that gives you an idea that how the genes are organised in the genome and of course there are large number of you know links that are there you can go on looking to for example what genome addition you are looking at, what cell type you are looking at and so on. Experimental matrix one can go on look into that.

(Refer Slide Time: 20:51)



I will show you one such snapshot of the genome browser, this is I going to show you a region of chromosome that is chromosome 6Q24 this is a gene that our group has been working on for a long time. This is called as CPM I just showed you to appreciate what you can get out of the genome browser. So it clearly tells that this you can see here there is a red line that is the region in which the gene is localised. You can see this is chromosome 6 right, this is short arm of the chromosome and this is the long arm of the chromosome and this is the position what you call as a 246Q24 that is where the gene is localised and you can see all these things.

So what does it mean so you see that there are lines here these are the regions that represent exons. So this is the five (())(21:33) of the gene and this is the three (()) (21:36) of the gene. Exon 1,2,3 and 4 so it clearly tells you how many exons are there and it not only tells you how many exons are there but it also tells you what is the distance between each exon. So you can go on look into these are the nuclear tide bases you can calculate exactly the distance and the span of the gene and in addition it is also telling you how many variants of the transcripter and you can see that you know there are you know so many trans 1,2,3, that start from here to here but you see there are differences for example this transcript starts here and ends here and there are 1 transcript you have 4 exons but in between there is an intra.

So how many different splice variants exists even that that are represented you can see here that is that this for the human and then it also talk about the what you call as methanation pattern of the chromatic. So what you can see here is the you can see like mountains these are segments

that are being shown to have methylated chromatic, so that is you know (())(22:41) for example astellation is show here right and it also talks about for example transcription you can see that these are how many transcript maps two different regions.

You can see there are mountains where ever you have exons obviously I going to have more transcript array sequence representing the exons and this one talks about chip data we will talk about a little later, what are the regions on which some transcription factors come and bind and you can see that there are transcription factors that are binding close to the three primer end but majority you can find at the five (())(23:16) end of the gene. Obviously because if these are the factors that regulate the genes expression they need to bind to other 5 (())(23:24) ends beyond that you can see that conservation for example if you look into the sequence and then look at the analogous regions are homologous region in other species how much is the sequence conservation right.

You can see that where ever you have for example exons the sequence conservation pretty high. You can see that even in chicken and zebra fish you find that the sequence homologous will exists right where as in intron then you do not obviously see much of the sequence similarity because these sequence are not under selection pressure. They can undergo changes because they does not affect the gene function but there are changes happening in the coding sequence it would affect the gene functions therefore such changes are not altered.

Now that talks about conservation but we come at the bottom of the screen and these are the hips that each bar represent regions in which you have variations in the human genome remember I told you about 1000 genome project and such project really talks about what is the variation at the genome level across the population and you can see that there are large number of sites you know in the genome that show multiple variants.

Now whether these variants alter the gene whether the alterations affect the way the gene functions whether that changes in the gene function you know provides you any risk of developing any disease or other conditions. These are all something that one needs to investigate but certainly that tells you that there are variations and that could be informative.

The last bar that is shown in here is the repeat right so the genome as you know that as large number of repeats and the repeats helps in the evolution in different ways. If you have read in

books that you understand that genes have evolved via duplication because of the repeats because it repeats itself at times in non-homologous recombination and as a result you have domains that are duplicated are genes that are duplicated and that help in the evolution.

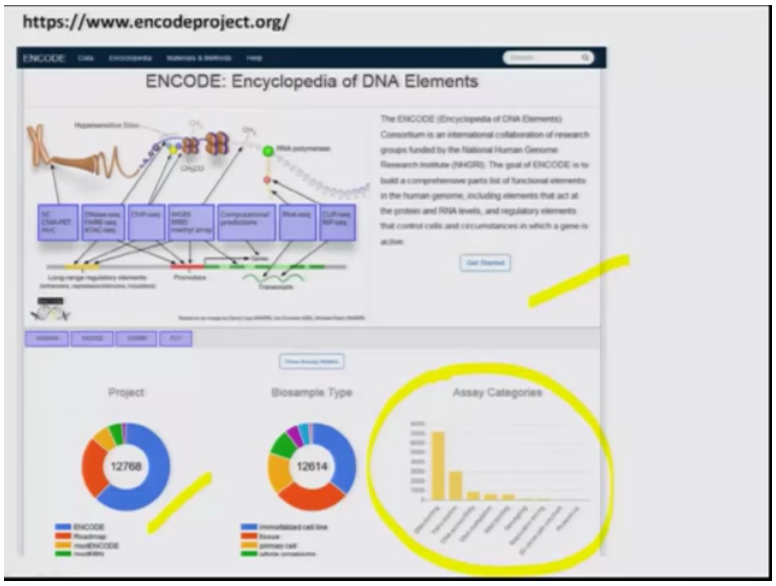
So what you see here is that there are repeats there are present across the gene and these repeats may have an influence or may not an influence on the gene function we never know but what could perhaps you may want to understand is that repeats at times can make genome unstable as I said the repeats you know at times can during recombination (())(25:52) when the homologous pair come together and then there is a recombination then instead of homologous aligning exactly at the same site because the repeats and their identical they can misalign.

As a result that could be an equal recombination resulting duplication of a small segment of the gene or a deletion of the small segment of the gene, so this leads to you know loss of function or gain of function and depending on how the gene affects the cell or the tissue. You may have one disease or the other, just to give an example this particular gene is involved in a variety of functions in the body and is very very critical for the neurons to survive.

So if you have some changes in the genes such that the gene is not functional then that individual would develop a condition called afro disease which is very very severe neuro degenerating disease. The person to begin with will be alright soon you would have problem in understanding memory, cannot walk you will have fixed what you call as epilepsy and invariably the person would die in about you know 25 years before that.

So that is really tells you how you know the genome sequencing and all other technologies that evolved including this kind of developing data bases helped us in one goal to understand the complexity of the genome and other functional signatures which help anybody who need not be working on the genome but can you know only working on a particular gene but it gives you every information which you can use to understand the gene function.

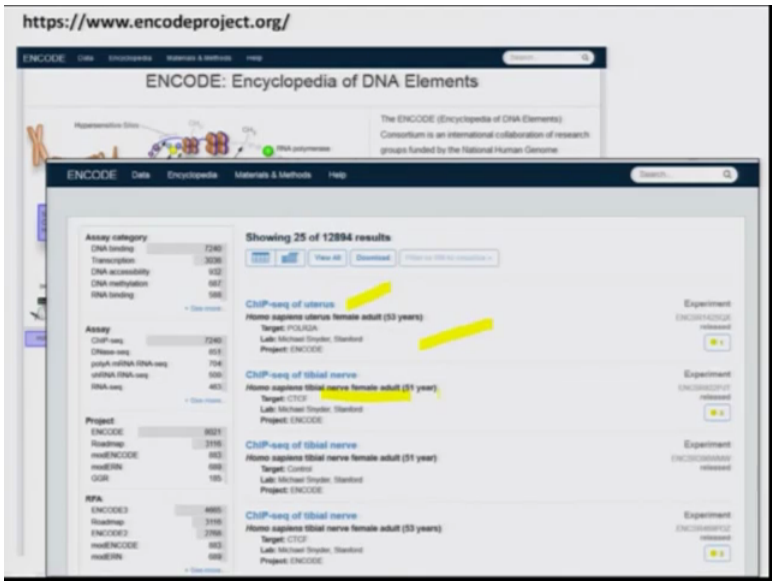
(Refer Slide Time: 27:33)



So just to give another example this encode project this is especially for people who are interesting in understanding the gene function, it is an encyclopaedia of DNA elements. It talks about data for example how for example what are the transcriptions factors that bind to the fiber and (())(27:48) of the gene and then for example the chromatin modification and then prediction, competition of prediction because you know with the genome sequence you will have a large number of data bases. Now there are data mining software people have developed tools to understand to predict genes or signatures and so on.

And this particular encode project really really helped us to put together and analyse and get essence out of the genome that is there and certainly you know you should go and have a look at it and that gives you more information. So it talks about for example the projects that are already there implemented and what is the road map and so on and it talks about you know all the essays that are you know put into this project, so it is good to go on look into right.

(Refer Slide Time: 29:02)



And you can basically look into for example the data, the data could be for example the chip sequencing data meaning chromate immune precipitation will talk about little later basically what you look into is that for example is there any transcription factor that goes and binds to different elements in the genome. So what are the reasons that bind to because transcription factors are the one that you know initiate transcription and in most of the cancers there is a global reprogramming in the way the genes are expressed.

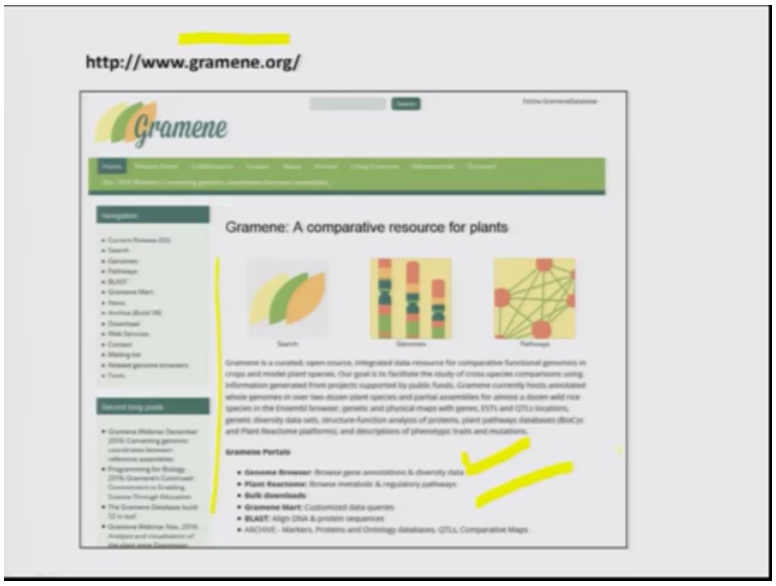
So if you can quickly look into what are the transcription factors and where are they go and bind you will know for example what are the genes that are up regulated or down regulated and for this people do what is called as you know chromatic immune precipitation and that data you know for example from uterus of female adult, you can see that and you know you can see it from the nervous tissue and so on. So you can get the data for each tissue of the human and either a normal tissue or from a cancerous cell, so that really you know you can go on mine more and then get more of understanding. So that is one snapshot of this particular encode project.

(Refer Slide Time: 29:54)

So the other application because of the genome sequence that has come is what is called as a vector based , in this again a data base dedicated to the genomic data of organisms that survives vectors say they themselves do not cause any disease but they can help in transmitting the parasites and other microbes to our body. As a result you know we end up having a disease for example mosquito this could be one for example malaria is another example and so on.

So you know this is again important area because you want to understand the host of the parasites which carry them and how you know their genome is organised and how do they function therefore the parasites are able to live there and reproduce because more often what we get infected is a certain development level form of a particular pathogenic right. So if you can understand that and if you can you know do something to their physiology probably we will be able to prevent or minimize the infection. So again its very very important there so that is something that there is come up and then it is not only the human and parasites.

(Refer Slide Time: 31:21)

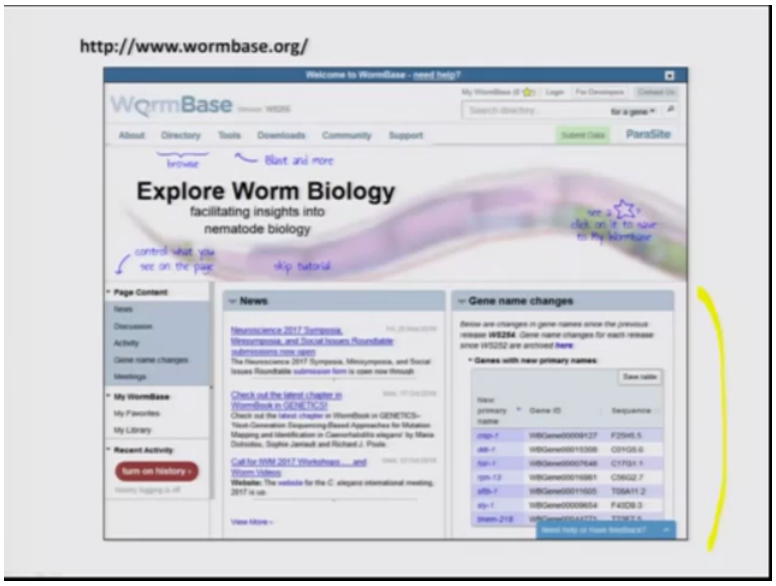


But even the plants are very very important for us because we are dependent on the plant for our food and these are the primary producers and which time you know that the land area that are being used for cultivation is shrinking because there is a urbanisation. So everywhere you see that there is a construction going on as a result you know the the land area that is used earlier for farming has reduced down.

Now what need to do the population is increasing, so you have a challenge with regard to how you are going to maximize the production of the food grains therefore you can feed the population but from a much smaller land area, so you need to make plant that are much more resistant to pathogens at the same time give more yield right at shorter duration right you know in the 40 days if you can get wheat out of plant so rice out of plant.

So you reduce the time required and and give more yield so that going to really help us, so to understand the plant you know the genome again there is a data base. You can go there is a Braim, data base again talks about all the genome that have been sequenced which you can go on looking to for example the genome browser which gives you the gene, annotation and the diverse data and and then of course the pathways involved in that and so on. So again people work on plant science and look into this data base to get more idea of it.

(Refer Slide Time: 32:47)



You have other data base there is a warm base like ways its stores all the information with regard to the (())(32:51) because is one of the powerful models people used to understand the development and the genes regulations and many other aspects. So this you know data base is got all the information with regard to how many genes are there, where they are expressed, what kind of function they confer and all the literature whatever comes the outcome is added to that therefore one can you know have a look at it. So that is again is going to be of great help if you go on look into.

(Refer Slide Time: 33:24)

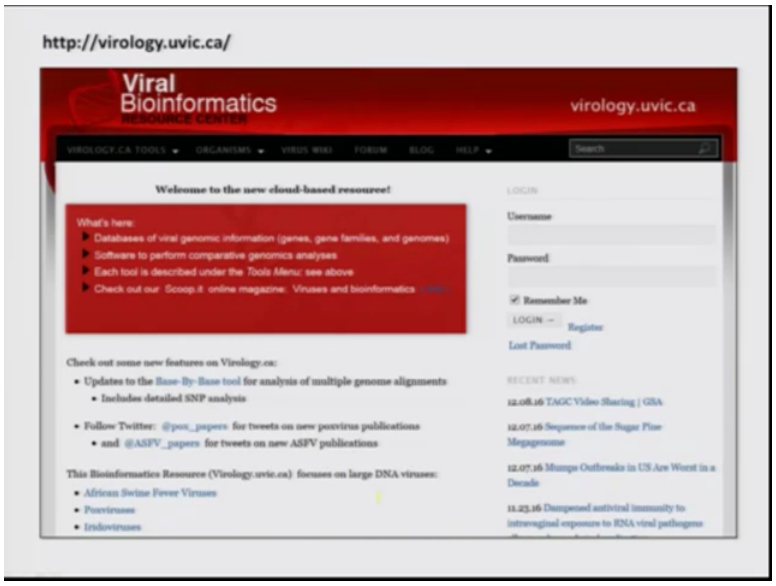
There is another model system FlyBase likewise it is a very good genetic model to understand the gene functions and development and it is not only talks about you know the drosophila genome and sequence but it also talks about you know the relationship between the genes that are conserved across species if you can recollect we discussed some time in last previous week about gene ontology. So the gene ontology is a data base is a dynamic data base which is being constantly updated with the known or the understood functions of you know various genes.

The gene could be studied in Fly but we have a you know similar gene in the human, so whatever has been studied in the Fly that data has getting to get into this data base, so even if you want to now study human that you know what has come from the Fly would help you in fact the pathways that are involved in cancer for example you know what has been tested in you know drosophila it seems to be very similar what has been found later in the humans.

So one of the ways by which people are trying to use the Fly is to mimic a cancer that are happened in the humans for example I have a cancer and my genome has been sequenced then that sequence information tells me that what set of genes that are altered in a cancer. Now the Fly being a very good model for genetics studies, so what one could do is within a month we can create all these combination that was found in the gene combination that were found in my cancer.

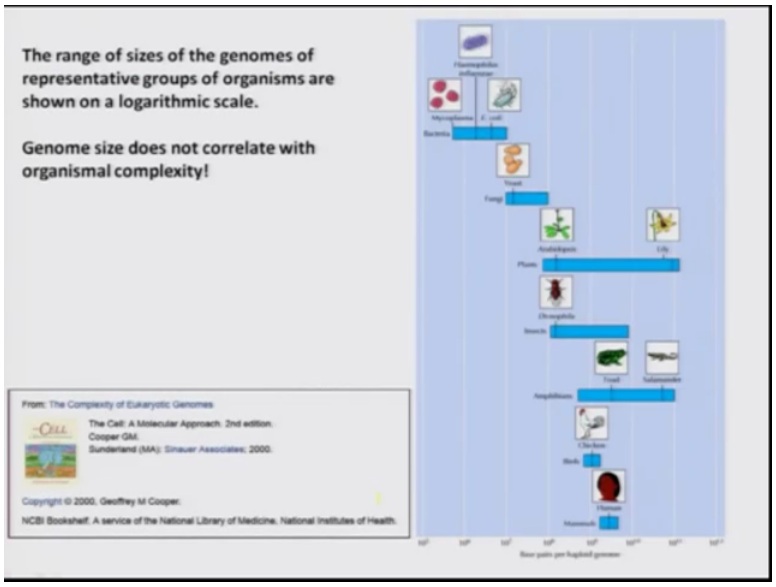
We can create that in the Fly model and quickly screen for drugs that are more effective in controlling this cancer and come back to me for treatment, so in this way I am going to give what is called as a personalised medicine. I am going to tailored the drug that are required to control my cancer depending on the kind of genetic operation that took place in my body. So you know the Fly is one such model that really helps to realise what is called as a personalise medicine. So that is where even you know studying an insect really helps in understanding what happens in the human and even to do a treatment, so that is the one good thing about all these model system and genomic information.

(Refer Slide Time: 35:44)



And finally we are going to show you one more such data base called as Viral Bioinformatics resource centre again this is data base that talks about all the viruses and their genome and their annotation and all the you know the publication that come out of this.

(Refer Slide Time: 36:02)



So that is the first lecture and we would end here and then the next lecture we will look into how we sequenced the genome right.