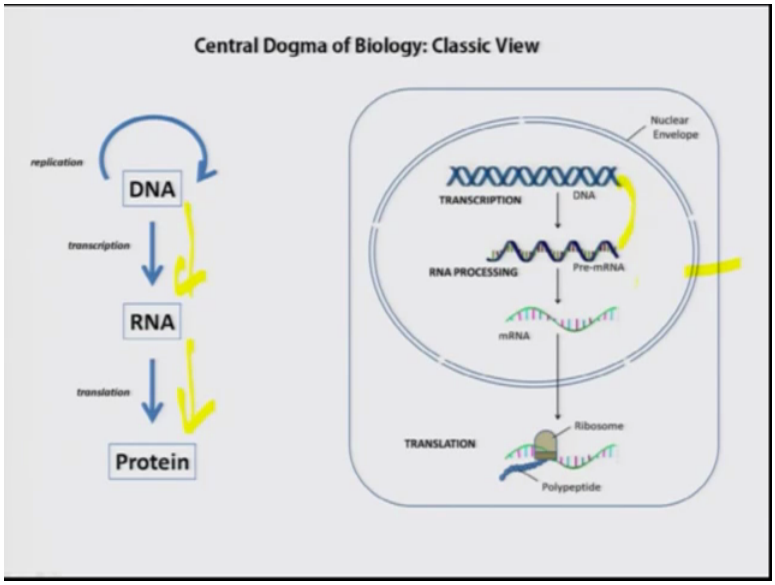
Functional Genomics
Professor S Ganesh
Department of Biological Sciences & Bioengineering
Indian Institute of Technology Kanpur
Lecture No 07
Transcriptomics Part 1

Welcome back to the course on functional genomics in the previous lectures we have seen how one could engineer the genome and and look into the function of a variety of genes in terms of a given process. We have looked at the example of cell division mitosis we have looked at the examples of virus virus infection process and then some of the developmental disorders people have moral dozing or identified the genes using zebra fish and so on.

So what we are gonna do in this lecture and the few lecture that are following this one is to understand what are the tools contemporary genomics tools people use to understand the expressions of the genes basically we going to look into the genome wired expression studies right. We look into the history and the tool how they are developed to understand what are the genes that are expressed for example you are looking at two different source tissue 1 and tissue 2 or it could be a normal condition and disease condition and what is the difference in terms of expression when you talk about expression it is not where are not a given gene expressed.

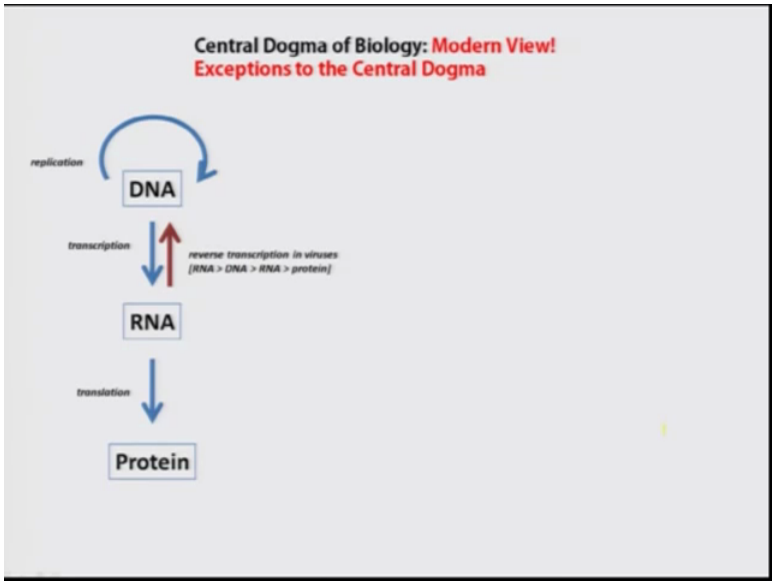
We are although looking at the change in the expression level you know the amount of the MRA that is being made for a given change, so that also dictates because higher the protein lower the protein it can change the way the cell behaves therefore we need to have methods to even measure the level of expression meaning the number of transcript that are being made. So before we get in to let us look into some of the basic concepts that we are going into what is called as Central Dogma of Biology.

(Refer Slide Time: 2:01)



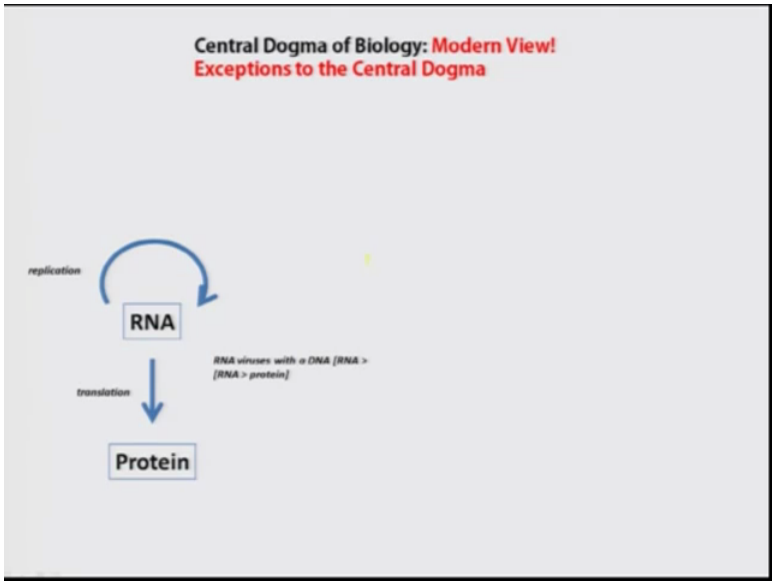
The classic view, the classic view is that DNA replicates makes copies of itself that is how it is able to divided and sustain and then there is a process called as a transcription which makes the RNA. The RNA is translated into protein that is what shown here in terms of inequalitic serve. We have the nucleus we have the DNA makes RNA and process to form messenger RNA and that is being translated to form the peptide or the functional protein.

(Refer Slide Time: 2:29)



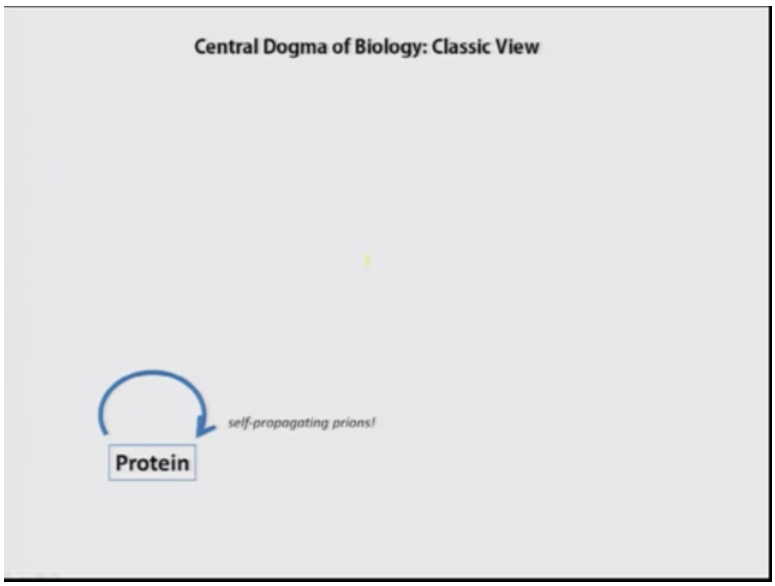
There are exceptions providing most of you are aware of the exception is what is called as the modern view of Central Dogma that is that that we discussed in the previous classes lecture that that you also have a reverse process what you call the reverse transcript which converts an RNA into DNA.

(Refer Slide Time: 2:50)



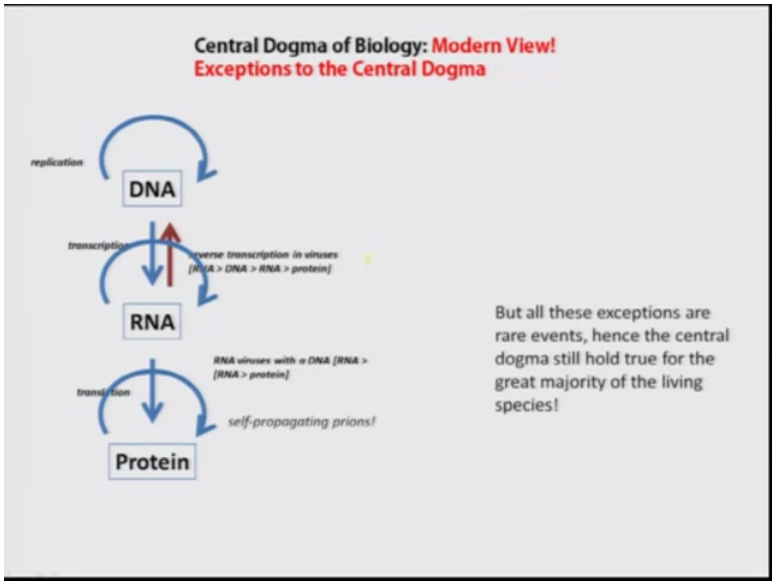
And that there are viruses that exists in RNA form, RNA genome.

(Refer Slide Time: 2:56)



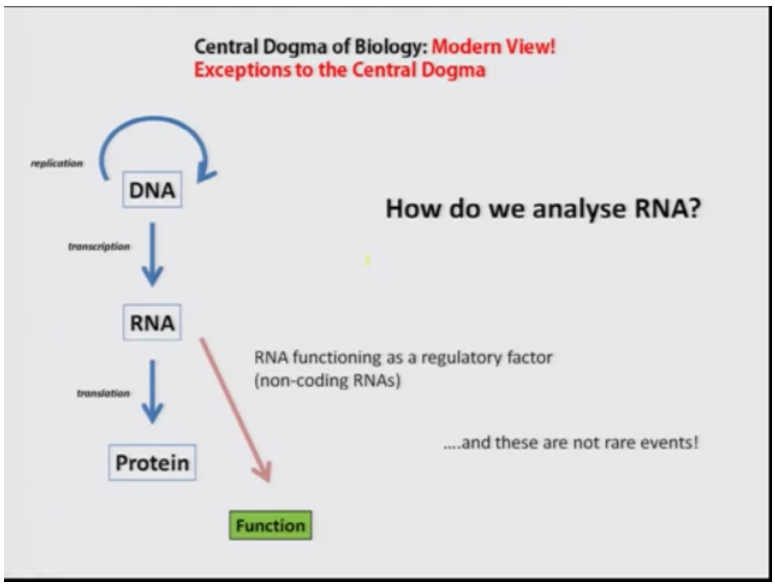
And there are also proteins that are self-propagating meaning you know they can make their own copies and so on.

(Refer Slide Time: 3:02)



So there are exceptions but all these exceptions are rare events but if you look into the majority of the species living entities than you will find that still the original proposal of Central Dogma Biology is valid and whatever the other exception that we said or exceptions.

(Refer Slide Time: 3:23)

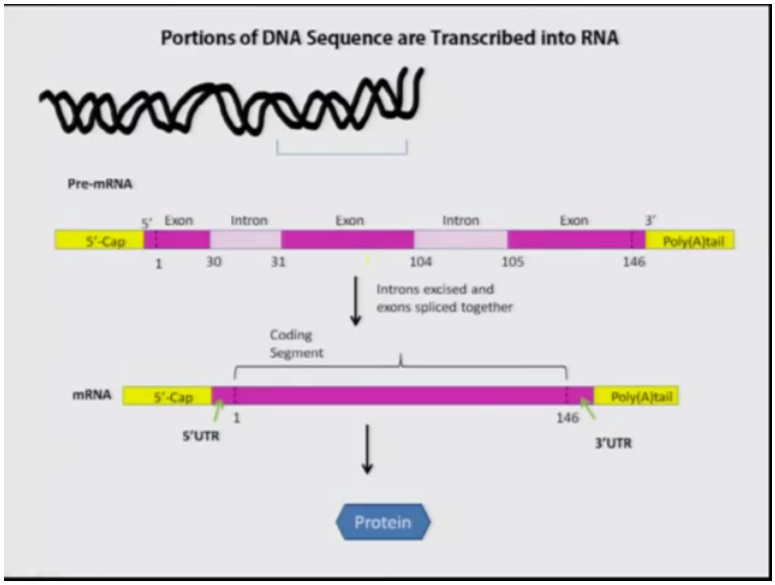


But what is important to know here is that it is not always that that the DNA is able to bring about a change in the cell system in the form of a protein via an RNA. In fact now you know that the majority of a transcript that are made in our cell including humans function as an RNA. These are called as a regulatory RNA as non-coding RNA in fact this fraction is much much larger than the RNA that codes for the protein. So these are called as a non-coding RNA that is the real real major deviation and as I told you these are not rare events.

Now people are trying to understand the function of this non-coding RNA. Some example we have seen in the previous lecture what is called as a microRNA or miRNA where you have discussed how this micro RNA is cleaved into smaller fragment and these fragments go and bind to some of the target RNA and destabilise them or prevent their translation and so on. So one such micro RNA can effectively regulate a large number of genes so that really shows that these non-coding RNA which do not really give the protein as a final product and not even helping in the translation process unable to regulate a variety of process.

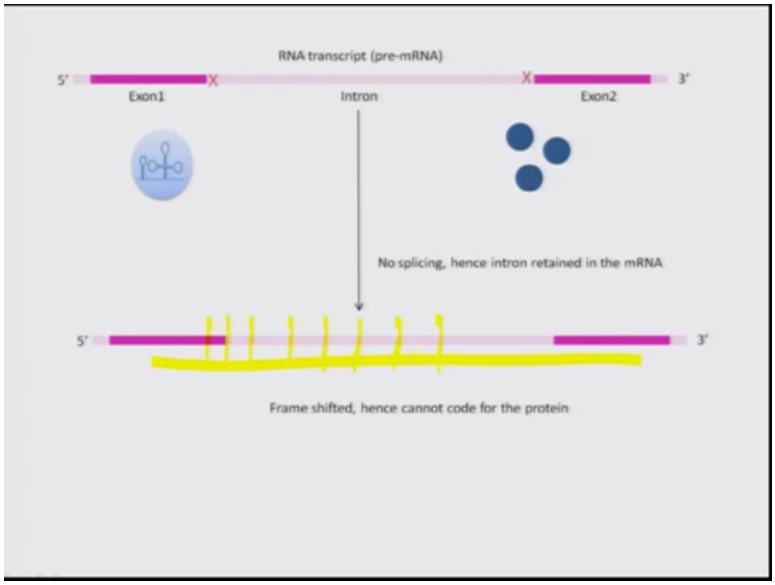
So they are these are not exceptional but these are very common that's what we believe. So how do we analyse these RNA that is a big question because analysing RNA is again tricky because you are going to look at the steady state because the RNA is being continuously made some RNA may not be translated they are made they are kept there some RNA are not making the protein but function as an non-coding RNA. So there is a huge challenge with regard to you know the analysis of RNA and secondly if you are looking at where the RNA is localized right. So you onto look into the localization of the RNA that becomes much more challenging.

(Refer Slide Time: 5:29)



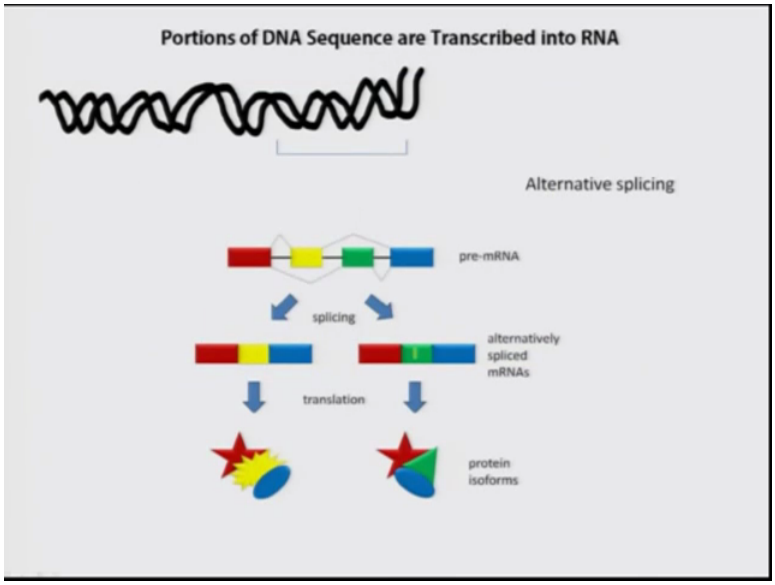
So let us look into other challenges the RNA pursue, the it is not that one gene makes one type of RNA right that that's not the case. Now we know that the RNAs undergo what is called as splicing. So what is shown here is schematic of a gene which as for example in turns here and here when this RNA is processed than final mature product you know the intronic regions are spliced out and we said that some of these introns can function as a (())(05:55) these are possibilities but you have the protein and off course we spoke about how the splicing happens and so on.

(Refer Slide Time: 6:08)



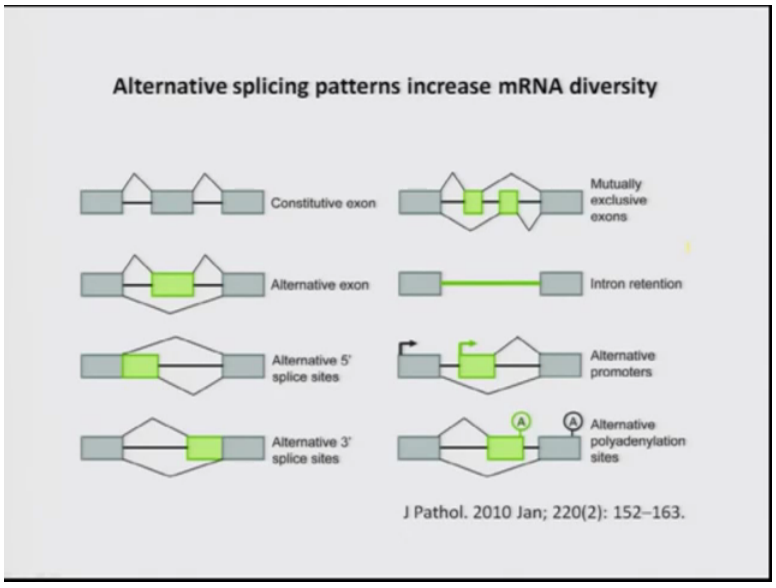
You know if if the splicing does not take place than you are going to have the introns retained in the in the RNA and these introns you know can introduce some change in the reading frame because you know you have triplet and if you have continued to you know make triplet because the introns are present here the reading frame will be shifted. So there probably going to have that but this you will normally call it as an aberration because the intron should be there and should have been removed and it is present now therefore exasperation.

(Refer Slide Time: 6:37)



But what you know is that such things are happening in the system so there are RNAs which have exons like is shown here 1, 2 and 4 four different exons and now they how this pre-mRNA having this four exons are assembled after over the splicing passes could be very. For example in this particular example you have exon1, 2 and 3 joined together whereas in other it is exon 1, 3 and 4 joined together as a result you know the coding sequence of this transcript and this transcript are going to be very different because this region that shown here in red colour may have a common coding region but beyond that the freeding frame will be shifted therefore this protein and this protein may look very different and may do very different kind of a function so possibility. So same gene which makes you know one pre-mRNA now can end up making different kinds of protein because a way the splicing takes place, so that is called as a auto dip splicing so therefore it is a huge challenge.

(Refer Slide Time: 7:51)

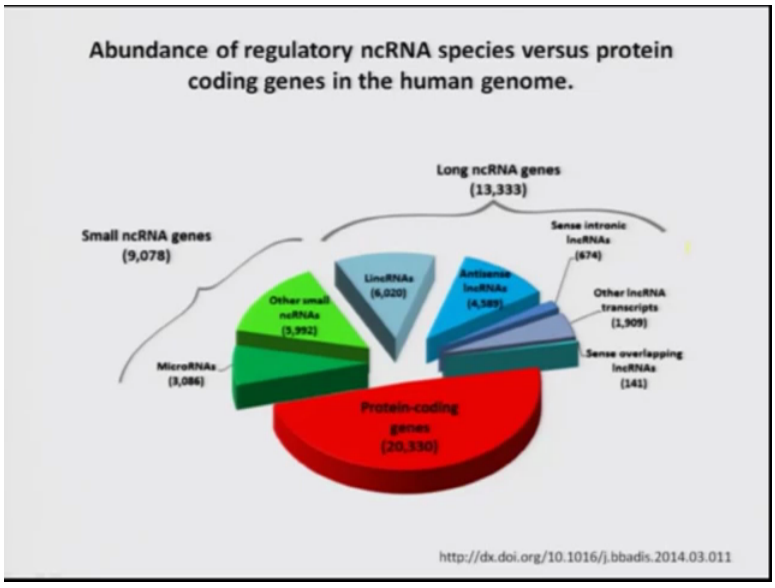


And there are a number of studies that are shown that a given gene depending on which tissue it is expressed which stage of the development is expressed it can make a variety of different splice warn and each one has a different coding potential and there is a relative level between these transcripts. For example what is shown here is a tow gene that course for one other protein that is called as tow which implicated in isomer condition. So it undergoes ordinary splicing and it is shown that in in aged brain that way it is spliced if it is modified right than the person may end

up developing dementia a process by which you know you lose whatever you have learned memory loss and and other symptoms that are associated with (())(08:39).

So therefore it is extremely challenging so it is not when you are talking about transcriptional analysis meaning an aluminate messenger RNA expression it is not simply whether the gene is expressed or not expressed its not simply whether how much the transcript is expressed but is also which transcript is expressed because there are variety of splice variants so you cannot go with certain molecular weight because the same molecular weight RNA can have different splice variant so it is a challenging task so that is something that we addressed .

(Refer Slide Time: 9:16)



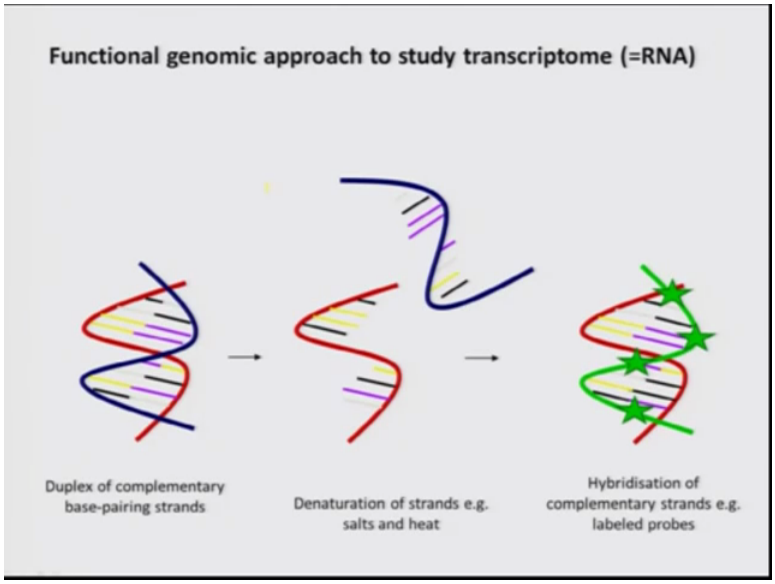
So if you look into the other challenges there not traditionally the messenger RNA you have what is called as non-coding RNAs and there are a variety of different types some are long some are short. So the approach that you are going to use for each one are going to be a different and you have to distinguish one group from the other because you cannot use the same group approach for same group.

So what is known is that you know you have this these are some of the classification that is come from recent paper, which says that you know you have this protein coding genes about 20,000 odd whatever it is than you have the other types these are called as small non-coding RNA genes because they code for what is called as micro RNA and so on which are very short RNAs.

We can see these are micro RNA other small non coding regions RNAs than you have what is called as a long non-coding RNA these are pretty long because it can run into several bases and there are other types for example you know intronic regions and other (())(10:17). It only bring shows you, the complexity that you faced when you want to analyse the transcript ok. So the answer is do we have all the approaches to identify all the different menu, types of RNA answer is no you know there are some of these RNAs the approach is yet to be fully developed.

We do have quite good number of tools to understand the expression levels of protein coding genes now we are having even approach to identify the micro RNA but as you discovered more and more RNA because even that is a challenging tool now. So what are the RNA that are really really the non-coding RNAs but having a regulatory function, so that you need to dissect first so it is an evolving field as is the case in case of function genomics so in all these conditions whether it is non-coding RNA whether it is micro RNA whether it is protein-protein RNA.

(Refer Slide Time: 11:34)

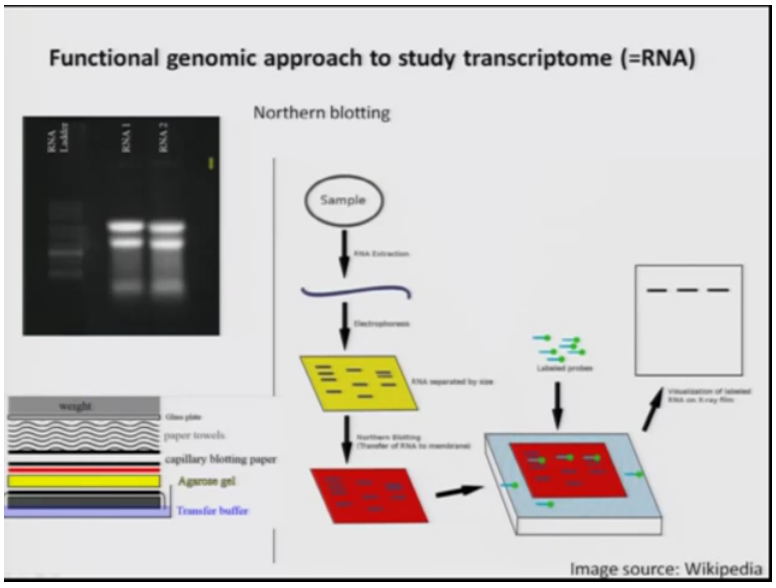


One of the fundamental tools people used to you know used to measure and quantify the RNA is the age old concept that is. Your RNA can form a duplex if you have a complementary sequence it could be RNA it could be a single stranded DNA but it does form a duplex. So if you have a complimentary sequence and if you are to measure whether RNA bound to back then you are able to measure the RNA so this is the most easiest way of looking at it and of course you can

always you know that that if you have a double helix and you the helix are held together via the hydrogen bonds that are found between the bases.

So when you destabilize them for example you can (())(11:59) by so example exposing to the heat or salt concentration than you are able to break the hydrogen bond and then we can again you know allow them to form the duplex. So if the the duplex you known one strand off course your RNA the other strand is something that you are making for a given change and that strand has some pro meaning ways to identify that for example you can add a fluorescent tag or you can put an isotope that is shown by the star here than you are able to visualize that. So this is the concept people have used since long to identify and understand quantify RNA.

(Refer Slide Time: 12:49)



So how do you do so let us look into the classic you know approach you know do not dramatically change this is the first thing that you do. You isolate the RNA. The total RNA from any species that could be human cells or animal or plant or whatever than what you see here the two bands these are separated in a gel so I am sure all of you know that the RNA can be separated according to the molecular weight if if the gel is run this way one of the topper larger species longer RNA the one that is below is a shorter species shorter RNA and when you put them across electric field in agro reserve you know the RNAs negative charge is going to migrate towards the positive pole and smaller the RNA faster they can migrate because they have force in the agro so this way you are able to size separate.

So what you see here the two band that are on the top is the ribosomal RNAs but you see here is the transfer RNA and what you see in between and you cannot see above is the messenger RNA because they vary in size whereas tRNA, ribosomal RNA the plant and small sub unit have different size of RNA there are multiple copies thousands of them so you are able to see like a band but the Rest RNA which represent them you know the messenger RNA is much less is 2 percent, 4 percent, 5 percent therefore and they are spread out different molecular weights so you see some of them here in between if you load if you apply more of the RNA will be able to see but you do not see here is the micro RNA which are very small it would have ran out of the gel so you would not see so you need a different approach further.

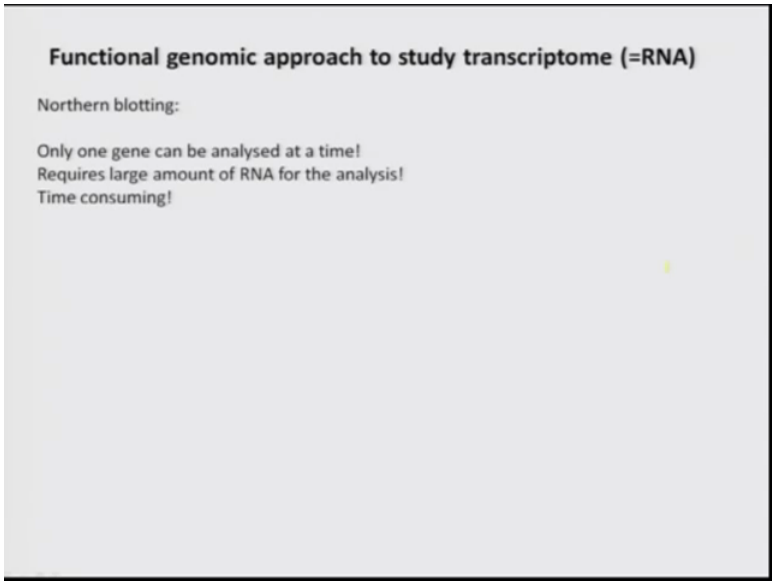
So this is a classic approach you separate them in a gel and then what you do separated by size and then you basically transfer the RNA that are there in the gel to a membrane like what is shown here you have the agro gel here and then you have the membrane you know and then you have this paper towels which are nothing but you know which has a capillary power you know pull the salt solution from here as the salt solution passed through the gel and the membrane it carries the RNA with that but it cannot you know the the RNA cannot go beyond the membrane its stuck there whereas you know we are able to transfer that. That is the way you are able to get the RNA separated RNA and the membrane and then you can put this membrane in a bag and use what is called as labelled approach.

These are nothing but a short stretch of sequences representing the gene of your interest say I am interested in gene X and I am taking a sequence single standard sequence of the gene X which is complimented to the RNA right. So this is what we have said all the approach that we are using is the duplex stability of the RNA to form a duplex. Now here we have labelled them because that is a way you can detect where the hybrid is where the bound RNA is present. For example you can label it with radio-active nuclear tide and then they go on bind because they form the perfect complementary sequence hydrogen bonding and then if you wash it so all those who are not bound are removed.

Where ever it is bound you are able to see if you put a x-ray film because than you have a radio activity that reduces and gives you that dark colour and there are other region where you do not have the radio activity. It looks clean so this way you are able to tell one whether the RNA is expressed two the intensity of the band whether it is very dark , light, moderate, that tells you

how much RNA relatively speaking is expressed. For example here you can see that the RNA level in all three samples are same right if this was much more intense than the other than you would say it is much higher as compared to sample 1 and sample 2 this is the way traditionally people used to quantify and measure un say whether RNA is expressed or not.

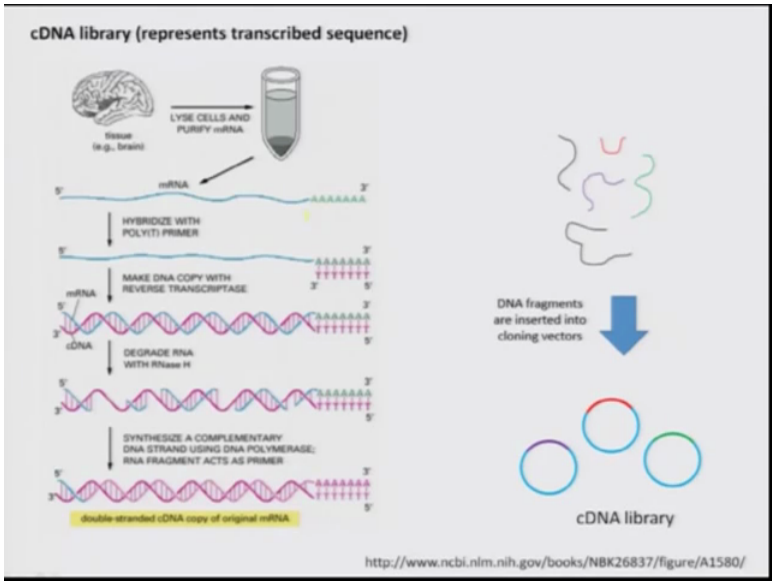
(Refer Slide Time: 17:07)



So this is called as a northern blotting because it was world after the first method for measuring or detecting the DNA which is then by southern blotting where in to separate the DNA the same way. Gel and gel and transfer and hybridize but here it was RNA that is why it is called as southern is called as Nolan, the first one was southern because that was developed by a person whose second name was southern and since that denotes the DNA transfer since the RNA transfer is different slightly so this called the northern blotting.

So it has advantages because simpler anybody can do it but as with any other approach it has its own disadvantages one is only one gene can be analysed at a time you have a block you can detect the expression of only one gene so after that again you have to make a blot and look for another and request large amount of you know RNA because we have to load high amount of RNA so if your sample size is small if you cannot get that much of RNA it becomes limiting factor. Its time consuming it takes two, two days to complete the process just to know one gene whether it is expressed or not expressed. So there are approaches people have used and we will look into some of them.

(Refer Slide Time: 17:55)



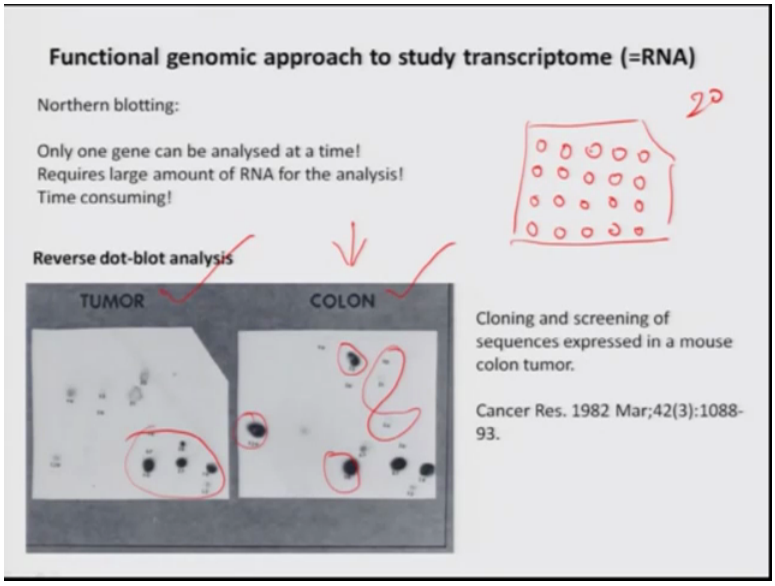
So one of them is to go for an approach using a DNA so what you do you can extract the RNA and use a same principle that is that you can have a small stretch of DNA sequence single stranded complementary to your RNA and which goes in and forms an hybridization like this so your most of your messenger RNA have poly(A)tail, so you have taken a tissue in this case for example the brain isolated the RNA and then you have selected the messenger RNAs and you have used this particular sequence the AAA poly(A) tail you have used an (())(18:36) which is complementary to this because you simply make multiple subties it go on bind and then you use a polymerase called a reverse transcriptase the one that we described sometime back.

So which copies the RNA into DNA ok it is become single stranded and then there are ways by which we can convert the single stranded DNA into a double standard DNA. So basically now we are going to generate DNA which represent the RNA so if you have had more copies of particular RNA we are going to have more copies of that particular cDNA. So it is not gene specific it is going to do it for every RNA that is present there which has this A so in other words this would represent pretty much what you have otherwise in the RNA.

So what do we do with this double stranded cDNA one can go on make libraries that is what you call it cDNA libraries but it also can be used for some other approach. First let us look into this cDNA library. So you have these RNAs which converted to cDNA which is double stranded now you can put them into vectors, theses vectors it could be plasmid it could be (())(19:46) whatever

it is but we can you know put these segments inside and these are you know made into cDNA libraries. So you have each one representing one particular transcript and stored in a library as and when you require you can take them and look into that.

(Refer Slide Time: 20:08)



In fact people have used this cDNA library to sort of increase your throughput meaning you want to look at expression of more than one gene but using very similar concept that used otherwise for northern blotting, how we can do. So this is one approach earliest approach which led to what you call now-a-days as microarray we will come to that a little later. What this particular group with this approach published in 1982 what they have done is.

They have selected genes or rather the cDNA of genes right which they were interested and then they you know took this DNA denatured it therefore it becomes single stranded and then spotted on a membrane, so basically what you are seen here is nothing but a membrane. So these are you know again the same nitro-cellulose are membrane right. So what you do is I have suppose 20 genes I have looked at, I want to look at their expression profile than what I am doing is I denature made into a single stranded form.

I am spotting it like this ok so this is way I spot it or each spot represent a given gene and all you have is a single stranded complementary sequence of the RNA that you expect that to be expressing. So what you do now you have the cells you know that that has got the transcription

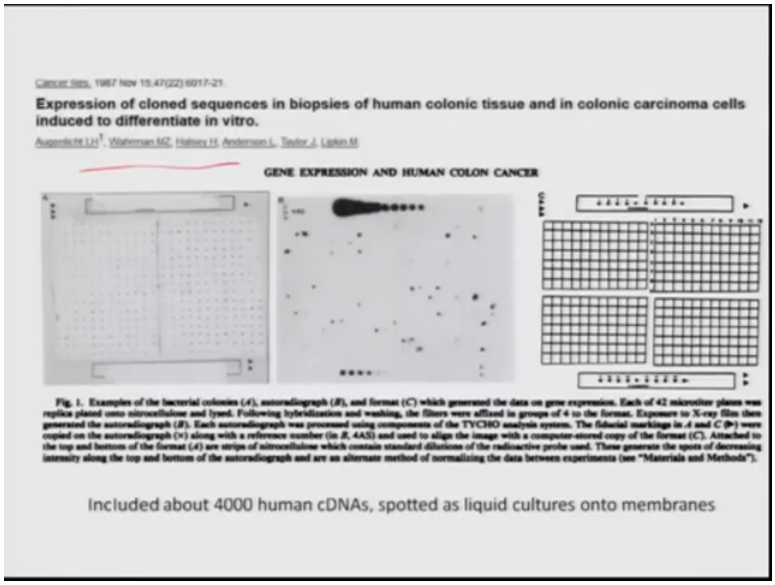
going on one way people earlier done is that in the transcription process you have added and labelled a nuclear type. So when the RNA is being made it has an isotope added to that right.

Now there are other ways you can convert that into you know RNA into cDNA in that process you can incorporate. For example labelled a nuclear type so this this the way to label the RNA or the cDNA that you make out of the RNA, then you hybridize that to the membrane say suppose this is a tumor, this is growing this s a normal tissue for example colan cancer I am looking at you know the normal for example right. So then I have the RNA you have the same blot same way I have a membrane in which I spotted the same number of genes.

Same amount of our DNA and then I am hybridizing with an RNA which is labelled now it all retried if I have a given gene expressing a high amount and therefore you would have more amount of that RNA and that RNA being labelled and if I have copies of that particular cDNA present on the membrane it will go on hybridize. So if I put an extra flame I may get something like this, these are the guards representing a given gene whose expression level is very high in this particular tissue for example colon you have this gene expressing very highly as compared to for example these genes they are you do not have much signal whereas here you have certain genes that are expressed very highly in tumor and some of them are express here as well but some of the which are expressed in colon are not expressed here.

So this is the way you are able to tell in more than one you know northern you are looking at one gene here you can go up to 20 genes 30 genes. We are able to spot them and tell the relative difference between the expression levels of the genes. So this is the very first approach people have used and then they went on to make much more densely spotted DNA not a membrane using machines now they gone into machine because even used machines to spot you know the tiny amount of a DNA on the membrane and this become commercial that where in fact companies were selling in early 80s and early 90s. They used to sell this kind of membrane spotted with you know 20,000 genes or 10,000 genes or whatever it is. Now that people have used to understand expression analysis

(Refer Slide Time: 23:59)



So that is what you see here this is nothing but a membrane and these spots represent a given genes a little bit of that complimentary sequence and then you have to off course isolate the RNA and label that and hybridize to the DNA spotted on the membrane and you can see that there are some spots that are coming up that shows that these are the genes that are you know expressing in a given tissue and the level and so on.

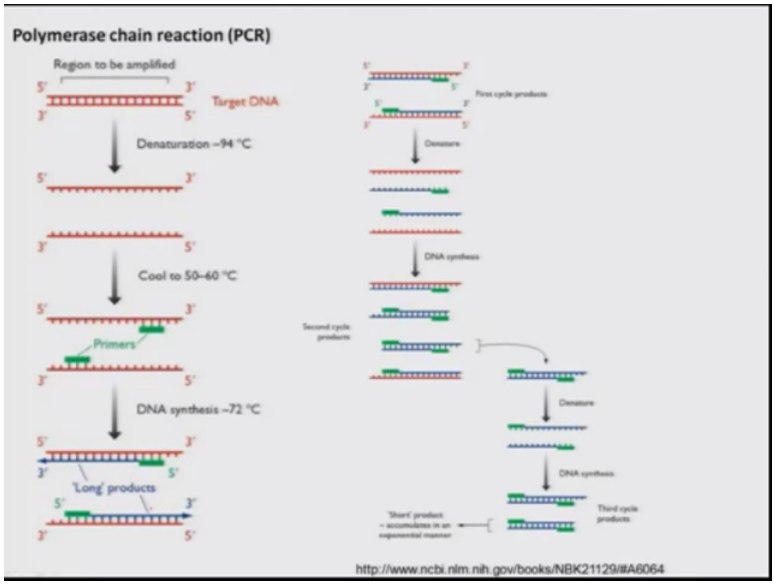
So in this we can put it over a gradient you know what gene it is right. So this is a way they came and in fact you know these there are about they could go up to you know 40 you know sorry 4,000 even cDNA there would have spot is you know it was a huge shift from the way people are using all and blot. You know that now people are looking at one gene at a time away from 100 you have gone up to 4,000 genes that was a huge shift by doing at this kind of approach.

So this basically leads to the next generation of approach which we now also call as a microarray right. Let us see what they are, the challenge now is that you have the cDNA clones right and the cDNA clones would have different size some cDNA would be shorter some cDNA would be longer. So when you want to really make an array like this ideally you would like to have two things one you would like to have the range of the DNA the size of the DNA may be 100 bases, 150 bases not beyond and so on.

Two, you want to keep the base composition to certain within certain limits. The reason being anything you hybridize you have to get rid of the hybrids that are not completely hybrids meaning there are some sequence matching, some sequence are not matching that shows that these are not the perfect duplex not representing the genes that you are looking at it. So you it how would you get it out that, you have to wash meaning you have to expose them to certain temperature which would de-stabilize any duplex that is not perfect ok.

So that you can do this temperature you can regulate only if you have allowed a duplex of base composition all can retain its complementarity in a particular window of temperature. It all depends on how many Gs are there, how many As are there right because it gives you that strength. So to all these things you should have some control over which region of the gene you are spotting to what is the length you are spotting, so these two cannot be done with cDNA clones because these are randomly made some are larger some are smaller. So it was very difficult to really control them so that is where they came up that's the time the PCR had rarely come something that most of you I am sure are aware of.

(Refer Slide Time: 26:49)

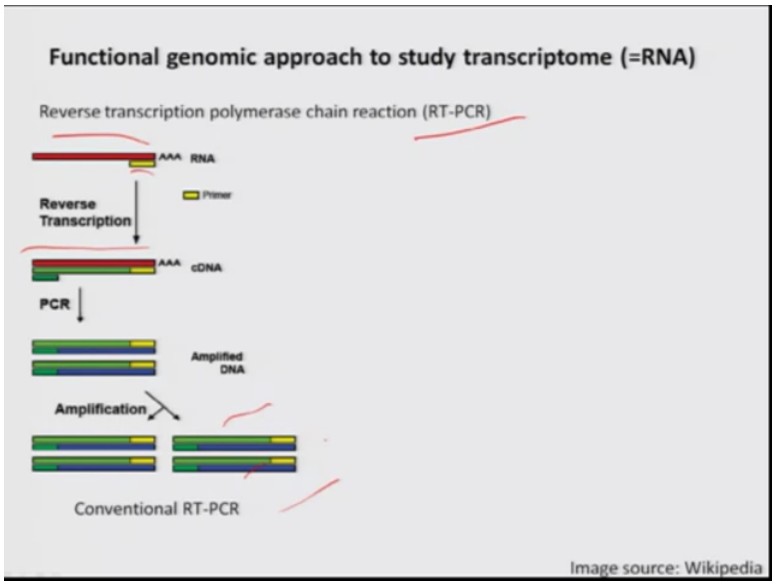


This is nothing but you have a DNA strand which you can make multiple copies of it by providing a pair of timers these are nothing but short allegro nuclear types complimentary to a given segment of the DNA. It goes and binds and then you have the so called DNA polymerise here it is something which can function at higher temperature not get de-stabilize by higher

temperature that is why it is called as a polymerase chain reaction because repeatedly are going to you know heat melt the DNA and then allow the primer to come and nil and then make copies of it.

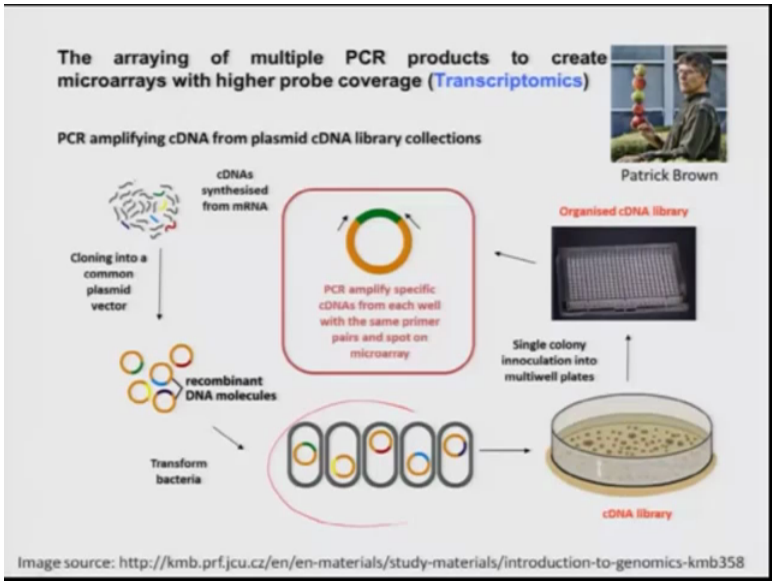
So this cycle we are going to make a short segment of the DNA, so this by designing primers for a variety of cDNA you were able to do achieve two things one the size of the cDNA that you are amplifying using PCR approach you have a control two given a base composition you will have control because you can choose a region that you want to make multiple copies.

(Refer Slide Time: 28:00)



This the way you know this you know the this PCR was done but if you instead of a DNA if you use a cDNA in your PCR than it is called as reverse transcription polymerase chain reaction or RT-PCR. So you have the RNA you have a primer which makes you know a copy of that RNA into what is called as a copy into DNA or cDNA and then off course you can make it double right. So by using this as a template in a PCR now instead off amplifying the genomic DNA you are amplifying the cDNA which represent the messenger RNA.

(Refer Slide Time: 28:27)



For example now what they have done is the following, so they have used you know these cDNA clones representing different genes using primers. They are able to make you know multiple copies of the cDNA and then make what is called as multiple plates to make what is called as the arrays that what come which led to the formation of microarrays. So these microarrays we call them micro because it has small parts which is microscopic array because it has got multiple spots ok and what it helps is that in a small membrane or a slide you are able to spot DNA representing thousands of genes so this cannot be done by hands like the previous one we have seen.

You will have try to sort of do it by hand up to some extent automation but this was much more challenging so they use a machines like what is shown here Automated robotic manufacture you know machines have come which are able to spot and people have also use what is called as bubble jet you know you must have seen bubble jet printers these are nothing but you know printers that have a spray kind of a technique. You know they just sprays that colour in a given region therefore you can get the image just like what you have you know.

So they have used even that kind of a approach and people have used even machines that used for the electronic circuits. You must have seen if you have open a computer you will find the circuits inside, these are nothing but you know these are printed on a chip. Therefore it is also

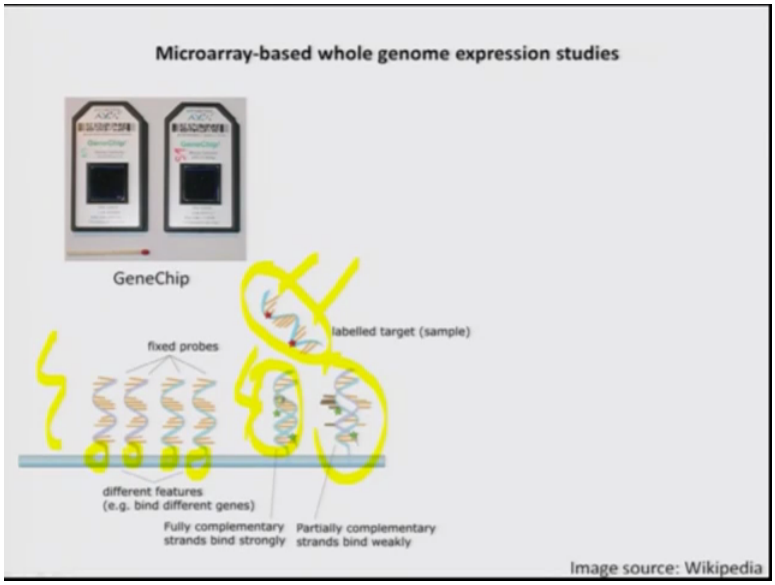
called as gene chip because what is being printed is a DNA over a chip using a very similar approach therefore is called as a gene chip.

So that is how you know these are some of the machines just show that how the DNA was spotted on a glass slide and then you came up with what is called as a high density arrays you know on a microscopic slides. You are able to now spot thousands of them and that really let to what is called as you know a chip having thousands of genes and now you are looking at expression of you know 10,000 genes, 20,000 genes representing the entire geno.

(Refer Slide Time: 30:43)

So these are some of those setups people have used to spot the you know array of you see that it is going more like a industry you know as I told you. Genome makes something like more like a industry because less of a human interference most of the mart taking care by the machines.

(Refer Slide Time: 30:57)



And then just showing you some of the first or second generation chips that have come. This is from a company called affymetrix and what is shown here is a match stick just to show you the comparison that you have the chip each one having you know tens and thousands of genes spotted over here. This region there are spotted ok, so they represent if for example if you take two chip that pretty much covers the entire gene that are present or at least coding genes that are present in the human or mouse depending on the species you can take it.

So what is done so again these are nothing but the same concept that you have a DNA single stranded onto which something comes in blind so which forms a duplex and you are exploiting the pace paring the hydrogen bonds between the complementary sequence. So how these are DNAs are spotted basically you have fixed the one end of the DNA therefore the other you know the DNA is open for forming the complimentary sequence. So these are fixed each spot now we have multiple copies of the DNA representing a gene.

So you will have thousands of spots each one representing the DNA right. So you have now the complimentary sequence which possibly represent your RNA our DNA will discuss a little later which are again labelled meaning you have incorporated some label by which we are able to now measure whether the there is a need duplex right.

So now in such labelled complementary sequence we will come and bind if there are perfect homology right because you are going to you know pass them through certain temperature therefore only the perfectly pad complex will stay there others will be washed out and you will now you are getting your any signal from this part that means the RNA is bound there. The how much is the signal that will tell you how many copies are present. So this is the way you are able to measure whether a given gene is expressed and what is expression level.