

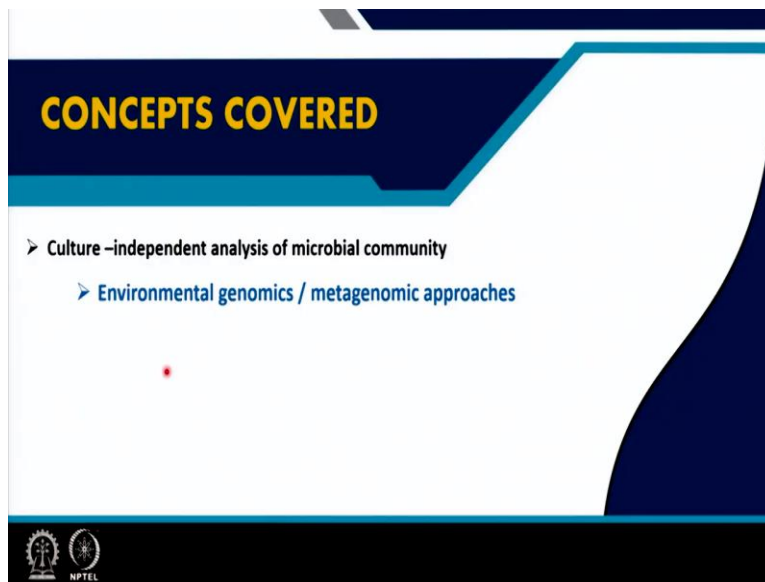
Environmental Biotechnology
Prof. Pinaki Sar
Department of Biotechnology
Indian Institute of Technology, Kharagpur

Lecture – 34

Methods in Microbial Ecology with Relevance to Environmental Biotechnology (Contd.,)

Welcome to the next lecture of this course environmental biotechnology and in this course we are going to complete our discussion on the methods in microbial ecology which are used in relevance to environmental biotechnology.

(Refer Slide Time: 00:43)



In today's lecture we are going to highlight the particular topic which is the the meta genomic approaches and this basically comes under the culture independent analysis methods that we have been discussing for some time now.

(Refer Slide Time: 01:05)

What is metagenomics ?

In Greek, meta : transcendent (beyond or above the range of normal or physical human experience)

Research field

Research techniques

The slide features a background with a stylized tree of icons representing various scientific fields. A small red dot is positioned to the left of the 'Research field' box. In the bottom right corner, there is a small inset video of a man with glasses speaking. The NPTEL logo is visible in the bottom left corner.

Now what is meta genomics? The word meta refers to transcendent that is something which is beyond or above the range of normal or physical human experience. Something which is large and something which is really able to provide; some comprehensive information or understanding on a particular topic or a particular subject. So, now with respect to this community analysis microbial community analysis the use of the term meta along with the genomics refers to some kind of protocol some kind of understanding.

Some kind of advancement that the scientists have made utilizing the concept of microbial genomes that means it is kind of an all inclusive concept that genome means collection of all the genes present in a particular environment and so, meta genomics basically has emerged as both research field as well as a research technique.

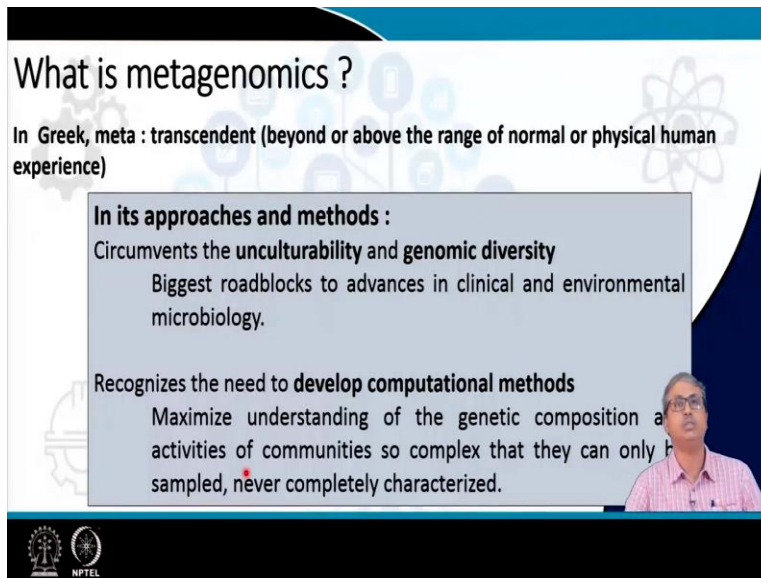
(Refer Slide Time: 02:25)


What is metagenomics ?

In Greek, meta : transcendent (beyond or above the range of normal or physical human experience)

In its approaches and methods :

- Circumvents the **unculturability** and **genomic diversity**
Biggest roadblocks to advances in clinical and environmental microbiology.
- Recognizes the need to **develop computational methods**
Maximize understanding of the genetic composition and activities of communities so complex that they can only be sampled, never completely characterized.





So, as a research field or a research technique we can see that it has actually diversified as I mentioned into two different directions or two different components. In one of the components that is basically targeting the methods and approaches it helps us to circumvent the unculturability and the genomic diversity. Now you are already aware of the unculturability aspect of any environmental community and the concept of the genomic diversity has emerged lately and what we have observed that within the genomes of individual species there are enormous amount of diversity.

For example if we are using 16s ribosomal RNA gene or we are isolating a couple of definite or specific group of bacteria and trying to look into the whole genome sequence of those isolated bacteria we are able to find out that within the members of the same genus or same species for example within E coli there may be different strains of E coli. Now if we look at the 16s ribosomal RNA gene these E coli strains might all appear same might they might not also sometimes they may show some variation at maybe 99% similarity or identity level.

But many a times we will be able to find out that all the strains of E coli or E coli by virtue of their 16s ribosomal RNA gene. But if we look at their genomes all the E coli strains that we have with us if we look at their genomes that means the complete genome sequence we will be able to find out that the genomes are significantly different. How significant this difference could be? This could be like 20% up to or more than 20% genes are different from one E coli to the other E

coli.

But all are E coli what that means to us who are working on environmental biotechnology it means a lot to us because suppose we identify there are certain E coli or there are certain vibrio or there are certain pseudomonas in our samples we cannot rely only on their taxonomic information that since there are pseudomonas or there are vibrio. So, they might be pathogenic. So, we have to take care about the quality of the treatment etcetera.

But it may appear that the genomes of those vibrio are having significant dissimilarity with genomes of the pathogenic vibrio and essentially we may not implement such kind of treatment processes if we even have those organism it is more appropriate when particularly we work on certain specific organism based bio remediation, bioenergy or sequestration of carbon dioxide or removal of phosphate or removal of nutrients etcetera.

We cannot just rely on the taxonomic identity of one organism. So, that means we need to have some technique which will go beyond 16s ribosomal RNA gene and will help us to take into account this genomic diversity that is the diversity that that is there within the genomes of even the same species or same strain of the particular species. So, now in this respect this new technique that is called the meta genomics helps us to overcome surely the unculturability because it is a cultivation independent method that means we work on the extracted nucleic acid directly.

But it also helps us to address the issues of genomic diversity that means if i want to know that how many different types of E coli are there in my sample although they may be similar or identical with respect to their 16s RNA gene still I can find out that. Now this particular aspect these two aspects the unculturability and genomic diversity these two were the considered to be largest roadblocks to advances in the clinical and environmental microbiology because all our processes all our diagnostic and treatment systems our ability to control the process etcetera were relied for a long period of time on certain type strengths or reference organisms.

We were unable to know that actually what are the other possibilities with respect to the; that

particular strain but having a different genomic abilities and the genomic ability is surely connected to their metabolic functions. So, the metabolic functions remain uncharted and also unexplored since the genomic diversity and the genomes were unexplored. Now it is also true that we cannot isolate all these species of E coli for example or all the species of burkholderia or all the species of vibrio.

But is there any way that we can still analyze the genomes of all those vibrio all those E coli all those pseudomonas whatever may be the genus or strain in my sample. So, that I can take appropriate decision that this sample has this kind of genomic diversity genomic potential. So, this should be utilized or this should be handled in an appropriate way. So that those road blocks unculturability road blocks and genomic diversity road blocks where actually at least significantly removed.

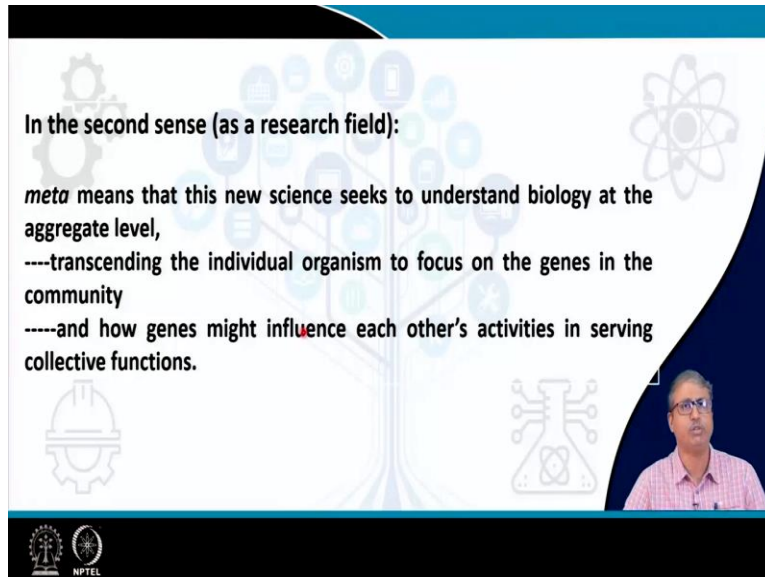
Now it also this meta genomics as a method or as approach recognizes the need to develop computational methods. Because we are going to rely on nucleotide sequence data and it is obviously not one or two sequence reads that any individual can handle manually. So, it is going to be huge and when we try to align the reads one or two reads aligning individually or manually is possible. Few tense are also possible but when we think of few lakhs or few millions reads it is not possible.

So, we are developing computational methods. So, or already developed a number of computational method in order to analyze this data is basically here the nucleotide or DNA reads that we generate during the next generation sequencing technique. So, that actually the developmental development of the computational methods that helps actually to understand the improve our understanding of the genetic composition of the activities of the communities and they are so, complex that they can only be sampled never completely characterized.

Because we cannot have the absolute value that what is the 100% of that community actually. We use different methods like 16s methods like earlier used to have the culturable methods. So, we are only deriving some information based on our capability but these computational methods are enabling us to achieve to some extent close to the absolute number absolute parameters those

are there.

(Refer Slide Time: 10:12)



In the second sense (as a research field):

meta means that this new science seeks to understand biology at the aggregate level,
----transcending the individual organism to focus on the genes in the community
----and how genes might influence each other's activities in serving collective functions.

The slide features a background with various scientific icons: gears, a tree with nodes, a DNA helix, a microscope, and a flask. A small inset video of a man speaking is visible in the bottom right corner. The NPTEL logo is at the bottom left.

So, in the second sense; so, that first sense was the method and the approaches and the second sense the meta genomics is basically in a research field it means that this new science seeks to understand biology at the aggregate level. So, it is something which is maybe I can use a corollary term which is called systems level understanding. So, it is not about one species one organism one particular gene something beyond that.

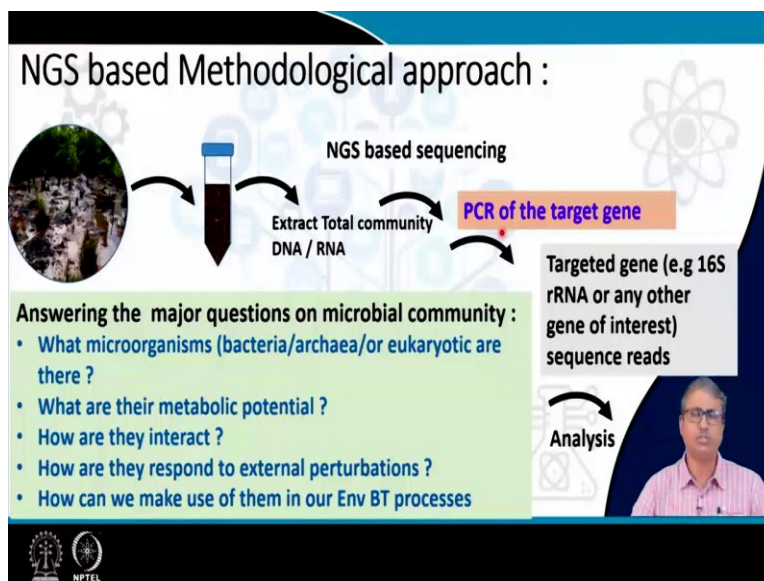
And by this time you must have understood that in real environment microorganisms are not alone they are. So, many they alone the prokaryotic microorganisms like bacteria and archaea they alone interact and they all the time they perform in within the guilds. So, they form metabolic links among themselves. So, they are all the time interacting with themselves. So, we cannot dissect one by one and identify a few and claim that we have understood certain wastewater treatment process or certain biogas plant system or something like that.

So, we need to have processes by which we can actually transcend the individual organism to focus on the genes in the community. So, instead of identifying one or catch-all of one or two bacterial genus or species we need to look into the genes which are there in the community. No matter the genes belong to whom if the genes are there pos the possibility is high that the genes are going to function.

And how these genes might influence each others activities in serving collective function that would also require that we have some methods through which not only the all the genes and their regulatory modules. Like the when we look at the open reading frame we can identify the regulatory modules. But also how these genes might be controlled in terms of the different environmental factors changing conditions it may be the changing conditions in a wastewater treatment plant it may be the changing because countries like India we have so, much of climate change like the heavy monsoon sometimes it is a dry spell.

So, the quality of our waste water the quality of the river or the lake water because of this kind of natural fluctuations they are changing all the time. So, how our drinking water treatment plants will take care if we are able to comprehend that these are the factors which actually control the microorganisms which are present in the in the particular water it may be drinking water or it may be some other type of systems.

(Refer Slide Time: 12:41)



So, now so, long the NGS based methodological approaches that we have been studying it has basically some components which are very clear to us that from the sample we need to get the genetic soup that is the mixture of all the organisms DNA and RNA. And then perhaps in earlier experiments or earlier lectures we have discussed that a PCR of the target gene used to be a very very important event because the target gene was a kind of a key component to characterize the

the genetic soup or the characterize the community.

Suppose it is a wastewater treatment plant or a kind of a activated sludge or a microbially enhanced phosphorus recovery plant. So, we were all the time focusing on a set of organisms that we knew these are the organisms who are responsible for this process. So, a set of specific primers often it is the 16s ribosomal RNA gene or some other gene of interest functional genes and essentially we were analyzing the data in order to answer the major questions which are like what microorganisms are there what are their potential?

How are they interact? How are they responding to the external perturbations and how can we make use of these organisms or this knowledge that what are the organisms present in our environmental biotechnology process. But is it enough during our earlier lectures when we discussed about even the next generation sequencing based amplicon approaches where 16s ribosomal RNA gene or similar genes are used to generate high number of reads may be millions of reads and then analyze it.

We have identified that there are some specific limitations then we limit the most prominent limitation is it is a PCR based process. So, it is using PCR primers which we know that they are well optimized they are very standard primers but who knows the target gene might have some mutations even in the pressure primer binding site. And even if the gene is there the gene may not be amplifying because the gene is having some new version some modified version of the nucleotide sequence at the PCR primary binding site.

So, even if the gene is there. So, I can actually clear one particular water as saying there is no vibrio or no toxic compound producing strain or no pathogenic strain. But actually there might be some because the pressure primer was unable to identify bind that particular gene because of some mutation or some alteration and also the availability of the particular template DNA and all other issues were there. So, we discussed earlier all these things.

(Refer Slide Time: 15:25)

Metagenomics

The advent of NGS technologies has provided easy access to a separate approach that is not reliant on the sequencing of PCR amplicons, yet capable of providing the information termed metagenomics.

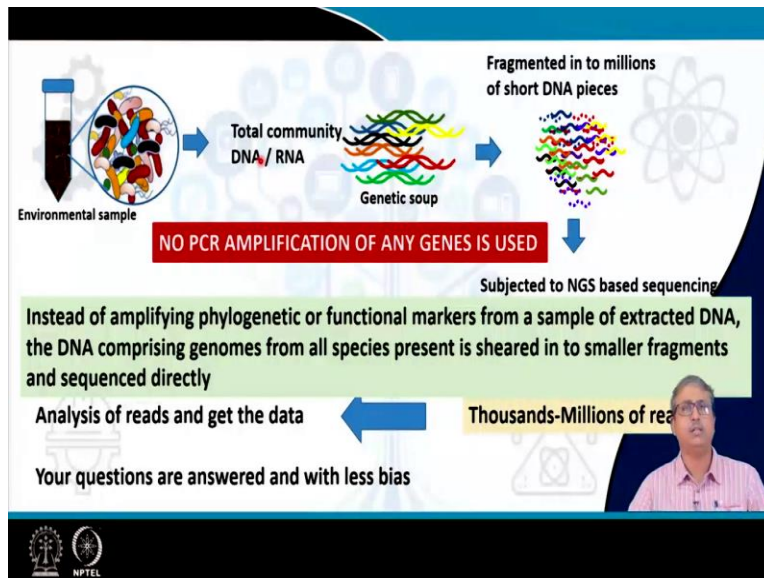
The slide features a central graphic of a tree where the branches and leaves are represented by various scientific icons such as a microscope, a beaker, a gear, and a DNA helix. In the bottom right corner, a small inset shows a man with glasses speaking. The bottom of the slide contains logos for NPTEL and other institutions.

So, basically the meta genomics when we if one when it evolves first it was maybe around 25 years ago or so and the last two decades over the two decades initially it was basically the fragmentation of the DNA of the total DNA without any PCR and the fragmented DNA pieces used to be cloned and the each of the clones is were screened for their sequence through Sanger sequencing or tested for their function.

Now the advent of the next generation sequencing technology has actually provided easy access to the separated approach. So, it is. Now not reliant on the sequencing of PCR amplicons for number one like we do not need to PCR anything neither we need to wait for the time and the money that large amount of large number of clones are to be sequenced because initially we were actually using or the scientists were using Sanger sequencing metho.

But the NGS method is capable of providing the information in a kind of a large scale which is basically meta genomics.

(Refer Slide Time: 16:34)



So, Now if I go ahead with this that we have the genetic soup that we discussed earlier. Now earlier if you remember we used to do a PCR reaction with this but now no PCR reaction the genetic soup is taken that is the extracted DNA or it may be RNA that I will come back come up sometime. So, total nucleic acid dot the DNA the DNA is straight away fragmented. Straight away fragmented using some methods which are chemical method or ultra sonication method and it is fragmented into small pieces.

Now each of these small pieces were subjected to next generation sequencing methods. So, that is the that is the beauty of the next generation sequencing methods that these methods are developed in such a way that they can handle a large number of truly large number of short reads which are generated and then larger number of reads are produced the reads are produced after the sequencing is done.

So, we get the nucleotide sequences which may be millions and sometimes even billions of rates. So, if we if we wish we can have billions of reads. So, what are these reads these reads are basically produced out after this sequencing of the sample. Now what was the sample? Sample was the short fragmented DNA no PCR it is the directly the DNA was fragmented and the small fragments obviously there will be some kind of a size selection process because the sequencing technology would prefer to have reads or the or the template DNA of almost similar size.

So, we will take enough starting DNA. So, that even if we have some smaller fragments we may need to sacrifice them but but we will still have representative DNA species and eventually we will get the reads those reads are going to be processed through the analytic analysis platform. Now these reads, now unlike the 16s ribosomal RNA gene or 18s ribosomal RNA gene or any kind of functional gene analysis that we discussed earlier these reads are not for any particular gene rather they are for all the genes those were there in the sample or those are there in the sample.

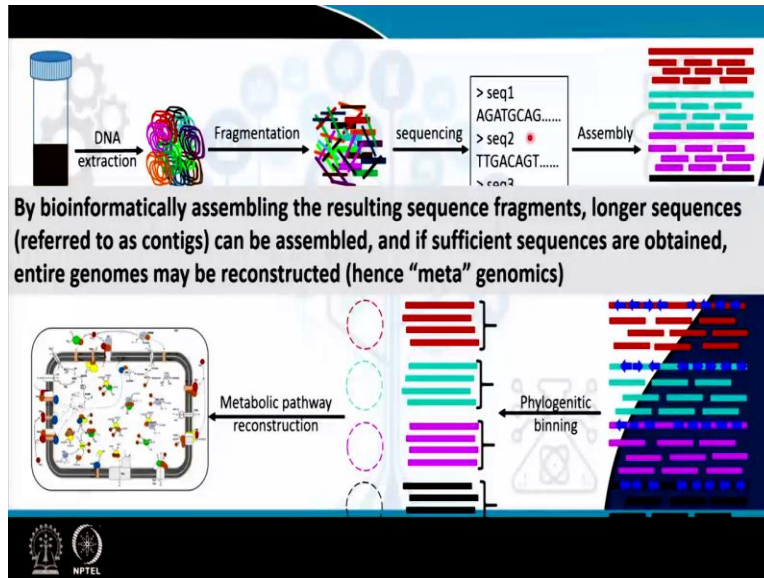
So, now when we when we try to analyze these reads we will get the information that will be entirely at the different level. So, that will obviously answer our questions but with less bias because PCR is not incorporated in this in this experiment or in this protocol. Now instead of amplifying the phylogenetic or functional markers from a sample of extracted DNA the DNA comprising genomes from all species present is shared into a smaller fragments and sequenced directly.

So, the this is basically having the intrinsic step that we extract the total DNA sometimes in some cases suppose for ground water or such sample we can actually have the cells sieved out first and then because we may have to use a large volume of water or particularly water maybe thousands of liters of water may be because the microbial load may be may be less. So, we may not have sufficient DNA at the end for the sequencing purposes.

So, we may need to pass a large by like 5 liter 10 liter water through the through the filter that which are very, very with very minor pore size 0.22 micron even less than that. So, that the cells are collected and then the cells are lysed no bacterial cells are isolated but the they are separated out in case of water particularly. But in case of solid sample like rock sediment etcetera or sludge kind of samples we generally do not need to extract the cells.

But again some scientists or some groups or some researcher might prefer to enrich and get their organisms isolated separately as a bulk mixed culture. So, that they can do but ideally the protocol may not require that.

(Refer Slide Time: 20:36)



Now simply as I was mentioning the DNA is fragmented fragmented DNA is subjected to sequencing. So, following sequencing we need to go for the assembly, assembly means the reads are joined together the short reads which are generated are assembled. So, for the assembly; there are certain logics used this assembly is done in such a way. So, that the pieces which perhaps belong to the same organism are joined together.

And the intact contig that is the contiguous sequence read is generated. Now once we have this read which are contiguous as you can see the reads are assembled to have these larger pieces of DNA they are subjected to phylogenetic binding that means their phylogenetic identity taxonomic identity is derived and based on the taxonomic identity we bin them. So, binning is optional some people like to bin it.

So, that you can identify how many organisms are, so, if you are doing a binning process then possibly you will be able to reconstruct the genomes out of the meta genomes but if you are not do not doing binning then still this context may be used for identifying the genes present there that is called the annotation.

(Refer Slide Time: 21:49)

Metagenome assembly

- Assembly is the process of combining sequence reads into contiguous stretches of DNA called contigs, based on sequence similarity between reads.
- Assembly merges metagenomic reads from the same genome into a single contiguous sequence (i.e., contig) and is useful for generating longer sequences, which can simplify bioinformatic analysis relative to unassembled short metagenomic reads.

The slide features a blue header with the title 'Metagenome assembly'. Below the title are two bullet points explaining the process. The background includes faint icons of a gear, a tree, and a flask. A small video inset in the bottom right corner shows a man with glasses and a pink shirt. At the bottom left, there are logos for NPTEL and a university emblem.

Now the meta genome assembly that I was referring to. So, assembly is the process of combining the sequence reads that is generated by the sequencer into contiguous stretch of DNA and these are called contigs. So, very short reads like 200 to 300 nucleotide reads generated are joined together. Now this joining is done by the bioinformatic tools which are based on certain assumptions. So, that the only the reads which are generated from identical type of genomes or the organisms basically.

So, they are going to be joined together to establish the or to get the contig. Now assembly merge merges the meta genomic reads metagenome derived reads from the same genome into a single contiguous sequence. Now the same genome is not visible to any bioinformatician or not even for the computer but there are certain logics which are used to to anticipate that these reads might be coming from the same genome.

It is basically some signature sequences or certain other residues which are identified and based on the bioinformatics program because there are multiple programs multiple pipelines are developed we are based on the type of samples there is some kind of requirement and that is useful for generating the longer sequences which can simplify the by the bioinformatic analysis related to unassembled short meta genomic needs.

So, it is always preferred that the short reads are assembled into context and then the contigs are

joined together to make larger fragments if you are getting a good quality data and able to use a proper bioinformatic pipelines perhaps you will be able to join the pieces like the contigs together to reassemble some genomes which will be called meta genome assembled genome or MAG ok. So, then you can actually characterize the entire genome of that MAG which is which is a kind of a reconstructed genome but from that you can actually develop further procedure and then verify them.

(Refer Slide Time: 23:41)

Metagenome assembly

➤ Assembly is the process of combining sequence reads into contiguous stretches of DNA called contigs, based on sequence similarity between reads.

➤ Assembly merges metagenomic reads from the same genome into a single contiguous sequence which can simplify bioinformatics.

Contig Assembly → Supercontig/Genome Assembly → Genome or Contig Coverage Profile

In some instances, complete or nearly complete genomes can be assembled, which provides insight into the genomic composition of uncultured organisms found in a community

The slide features a diagram showing the progression from 'Contig Assembly' (represented by several short red horizontal bars) to 'Supercontig/Genome Assembly' (represented by a single long red horizontal bar with a red circle below it), and finally to 'Genome or Contig Coverage Profile'. The slide also includes a small inset photo of a man in the bottom right corner and logos for IIT Bombay and NPTEL at the bottom.

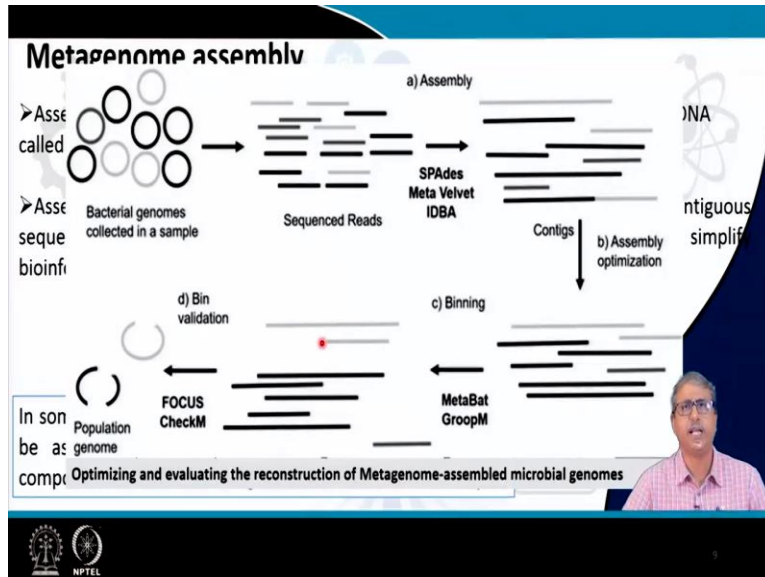
So, this is some graphic is here that we first go for the contig assembly and then super context or genome assemblies are prepared. And if you are having a very good data and you are using this appropriate bioinformatic pipeline then the shorter pieces joined into medium large pieces and the medium large pieces are joined to larger species and the larger pieces are joined to have the almost entirely complete genome.

Sometimes the genome may be incomplete because you may still identify there are certain gaps and some contamination may also be there but this is possible and it is a kind of a regular event for scientists who are using meta genome assembly technique to get the genomes out of the metagenome.

Now in some instances the complete or nearly complete genomes can be assembled which provides insight into the genomic composition of uncultured organism because otherwise these

organisms are not possible to be cultured. So, we do not get their genomes genome genomic sequence with us. So, it is only the meta genome analysis which help us to get the genome even if the organism is uncultured.

(Refer Slide Time: 25:01)



So, here again there are many methods through which this can be compared. So, here for example this pads and meta weld weight and IDBA these are different assembly pipeline winning pipeline these are the softwares or the bioinformatic validation pipeline etcetera. They are developed and used to have the genomes or even you have the contigs also. So, this contigs itself are good enough.


Because you with this contigs you can still go for your annotation that is identifying the genes that is there and then identify the organisms which are present identify the function that they can carry out.

(Refer Slide Time: 25:40)

Major outcome of the metagenomic analysis

Metagenomics answers four critical questions

1. Who is there ?	Community composition- All organisms
2. What they can do ?	Metabolic potential- All genes → Pathways
3. What are they doing	Realized metabolic potential: Select RNA to work
4. How do they interact with themselves and abiological environment ?	Network and models- Use multiple samples and datasets



So, what is the major outcome of the metagenomic analysis. So, the basically the meta genome is still able to answer the four fundamental question that is who is there that that means the the organisms present what they can do that is their potential what are they doing and finally how do they interact with themselves and a biological environment. So, in terms of who is there it is basically community composition that we get.

But unlike the 16s ribosomal RNA gene which is only targeting the prokaryotic organism in case of meta genomics it is beyond any boundary of prokaryote or eukaryote it will give you all the taxonomic marker genes 16s 18s even the RPO b gene and many other single copy marker genes which are conventionally used to delineate the taxonomy of the organisms both eukaryote prokaryote virus everything.

So, you will get the community composition beyond any taxonomic boundary what they can do if you remember earlier when we are trying to do this same answer in the same question in NGS based methods or other methods we are targeting genes specific genes. But in this case since it is relying on the total genomic pool it will give you the metabolic potential through all genes. All genes present in the environment that particular environment that particular sample the information about them will be in front of you and since you have all the genes or information about all the genes.

So, you can actually build the pathways, pathways which are relevant for carbon metabolism for nitrogen metabolism for sulphur metabolism for transport for any for a disease development for antibiotic resistant for fatty acid synthesis for carbon dioxide capture and sequestration into carbonate minerals for any kind of pathway. Basically you will be having opportunity to reconstruct all the pathways out of this you will select that which pathway you want to emphasize more provided of course if you have sufficient good quality data with you that is of course a point and I will come to that point.

What are they doing that is again the fundamental thing that the realized metabolic potential that means the function that they are actually carrying out. So, the first the question number two was what they can do that is the genes they have. Now whether the genes are actually getting expressed or not for that possible we need to select RNA and if we are taking the RNA out of the sample instead of a DNA then the total RNA that means it is the meta transcriptome.

So, the transcriptome metatranscriptomic sequencing in the same way would give us the realized metabolic potential all the genes which are being expressed or are being function, function being carried out. So, and finally how do they interact with themselves and a biological environment of course for that we need to have the network and models and we need to use multiple samples and multiple data sets.

(Refer Slide Time: 28:39)

The slide features a blue header with a gear icon on the left and a molecular structure icon on the right. The main title is 'Data on metagenomic profiles of activated sludge from a full-scale wastewater treatment plant'. Below the title are the authors' names: Jianhua Guo^{a,b,*}, Bing-Jie Ni^a, Xiaoyu Han^c, Xueming Chen^a, Philip Bond^a, Yongzhen Peng^b, and Zhiguo Yuan^a. A CrossMark logo is positioned to the right of the title. The affiliations are listed at the bottom: ^a Advanced Water Management Centre (AWMC), The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia; ^b Key Laboratory of Beijing for Water Quality Science and Water Environmental Recovery Engineering, Engineering Research Center of Beijing, Beijing University of Technology, Beijing 100124, PR China; and ^c Beijing Drainage Group Co., Ltd, Beijing 100022, PR China. A small video inset of a man speaking is in the bottom right corner. The footer contains logos for the University of Queensland and NPTEL.

Data on metagenomic profiles of activated sludge from a full-scale wastewater treatment plant

Jianhua Guo^{a,b,*}, Bing-Jie Ni^a, Xiaoyu Han^c, Xueming Chen^a, Philip Bond^a, Yongzhen Peng^b, Zhiguo Yuan^a

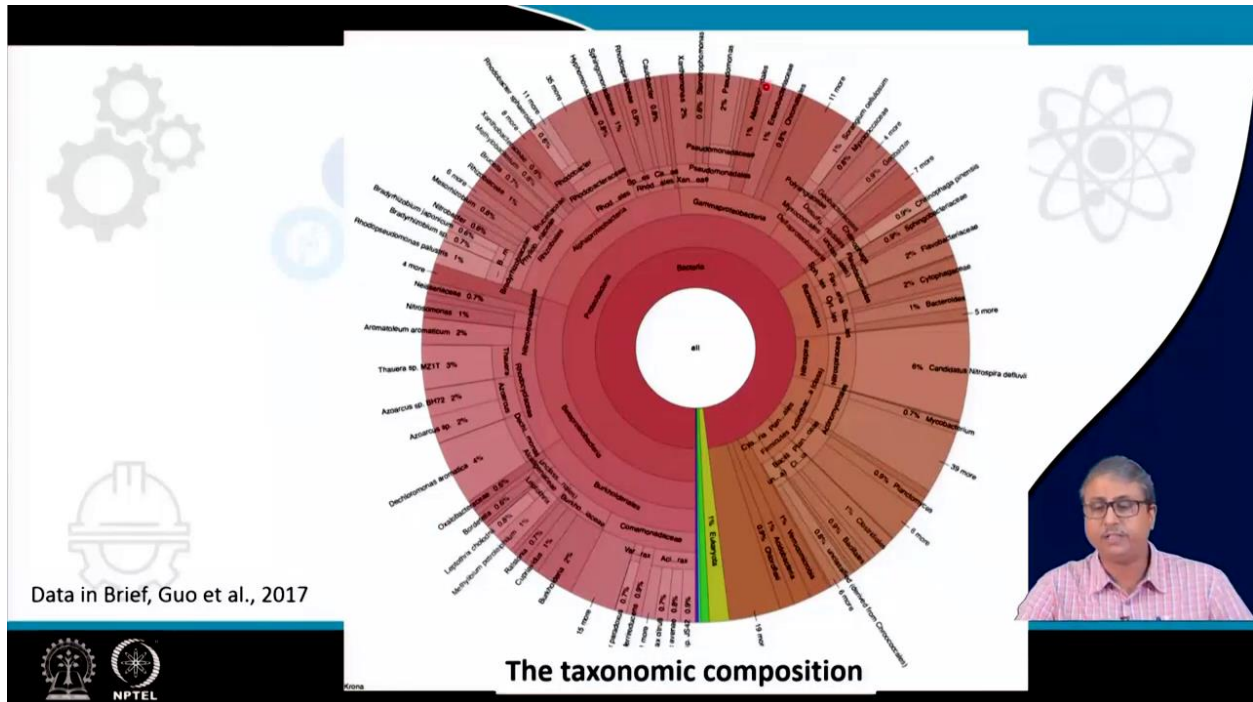
^a Advanced Water Management Centre (AWMC), The University of Queensland, St Lucia, Brisbane, QLD 4072, Australia

^b Key Laboratory of Beijing for Water Quality Science and Water Environmental Recovery Engineering, Engineering Research Center of Beijing, Beijing University of Technology, Beijing 100124, PR China

^c Beijing Drainage Group Co., Ltd, Beijing 100022, PR China

Now I will give you one example where you can see that the this is this is from the particular paper that is on the meta genomic profiles of activated sludge from a full scale waste water treatment plant. So, we can see that we have used this waste water treatment plant sample.

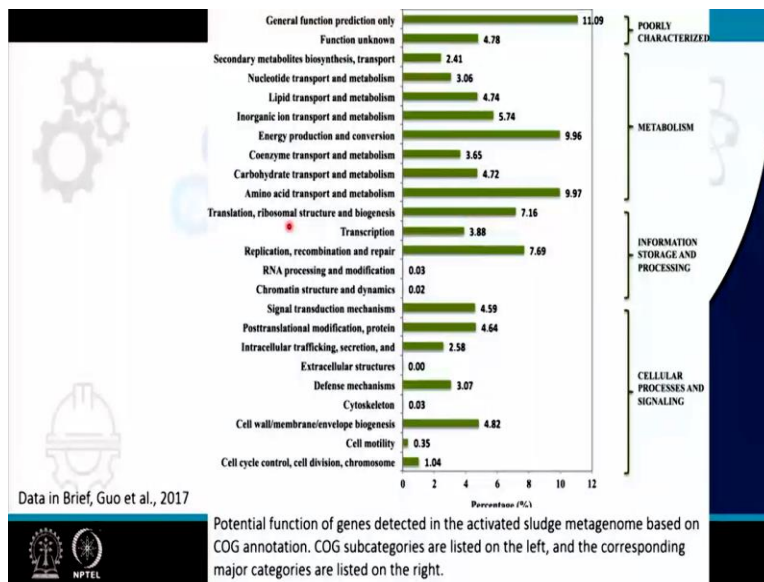
(Refer Slide Time: 28:59)



And when we perform the whole meta genome analysis of the scientists they perform this go at all they perform the whole meta genome taxonomic profiling incorporating all the taxonomic information you can see that. It is again dominated by mostly the bacteria and there are certain kind of eukarya and other organisms are there. And again it is mostly the proteobacteria and at organismic like genus level these are the data this is a chrono chart we call it and these are the all the organisms and their relative abundance which is plotted over here.

So, now you can compare this with indeed since the wastewater treatment plant with different samples or under different treatment conditions and then evaluate the composition.

(Refer Slide Time: 29:40)



This next is the function that is the very, very interesting thing. As I mentioned earlier the function here is delineated not based on any particular target any single gene but it is rather based on all the genes which are present and all the genes which are detected. So, there are certain systems through which actually we annotate. So, here it is the cluster of orthologous groups and notation cog and notation.

So, through this annotation when we try to annotate, annotate means we try to find out the meaning of those contigs those genes that are present there which type of reactions they carry out. So, generally these there are certain systems that we have identified within the genomes or the genomic property based on that. So, we in this case these are the major categories that are highlighted like the metabolism category information storage and processing category cellular processes and signaling category and some other categories are there which are called poorly characterized.

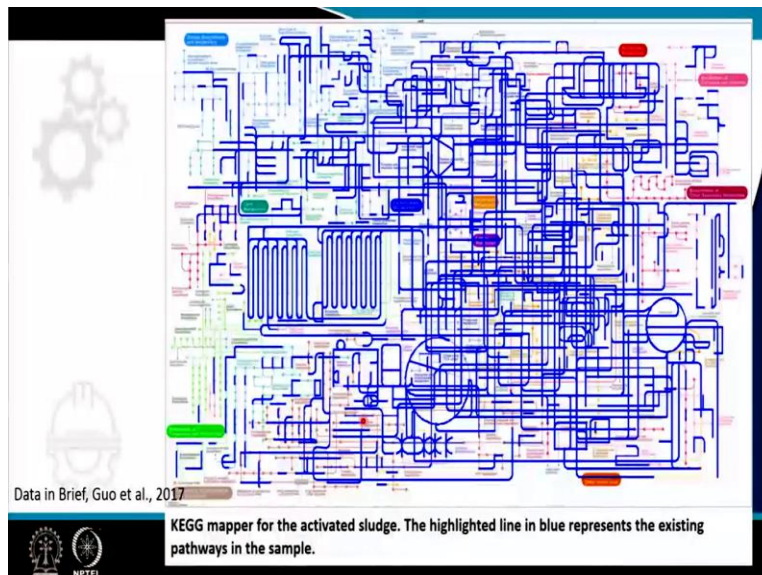
So, there might be other categories also but for the simplicity the figure is containing only this three and four the poorly characterized one because the group might have found some genes allocated to this. Now within each category like or the system the sub metabolism category it is a major category there are numbers of pathways. So, these are the pathways. So, these pathways are maybe certain some little higher level of pathway which are you can see these are called subsystems.

So, within a system of or category of metabolism there are sub categories which are like secondary metabolism nucleotide transport and metabolism lipid transport and metabolism inorganic ion transport energy production and conversion etcetera. Now energy conversion and production these has significance allocation of genes. So, these are the percent allocation. So, out of the total genes those are detected.

So, around 10 percent genes are allocated to energy production and conversion. Now if we want to look into this particular pathway or this particular sub category that what are the actual pathways which are going on in this particular activated sludge or the wastewater treatment sample which are facilitating the energy production. Then we can do that because we can find out that 10 pathways are going on and each of these pathways are having maybe 10 step or 20 step or maybe 5 steps.

So, all those genes are detected and their information are collected together collated together and they are put under the heading energy production and conversion similarly transcription replication and all things are done under the information storage and processing cellular signaling is also having many features.

(Refer Slide Time: 32:28)



So, like that eventually if we want to map the entire process happening in this particular sample

this is called the Kegg Mapper which can provide us that based on the genes which are detected and their relative abundance what are the processes metabolic processes are going on and what is the overall nature that we are talking about the systems level understanding. So, within the activated sludge or the wastewater sample there may be millions of bacteria of several 1000 species.

So, rather than relying on the functions carried out by one or two or three or four species if we can see that what is the total function carried out by this sample or the community of this particular waste water system. So, this is the picture over here from this if we if we resolve it because it is it is not possible to resolve this over here. But if we can resolve it in a live online system of KEG database will be able to see for example here if we go we will be see that the the site rate cycle of the TCA cycle is almost complete except this few steps.

Few steps the initial few steps are missing maybe the this is bypassed by some other reaction and somewhere here we have the glycolytic reactions and we have the other reactions of the energy metabolism and amino acid metabolism which are also very active. So, we can readily understand that at system level how this community is behaving. Now if we want to make use of this like if we want to produce or look into this community that whether we can produce certain useful molecule like some energy molecule some lipid molecule.

So, that we can actually convert this process into a part of the circular economy or something like that then we can actually fish out or we can actually look in detail of those genes which will allow us to identify that whether these genes are there and how that can be that can be helping us.

(Refer Slide Time: 34:27)

Key points of advantages

By examining the functional and metabolic attributes of the sequenced genomes it is possible to:

More accurate identification of the microorganisms present
Ability to predict the complete functional profile of a community

Metagenomic approaches avoid the known biases associated with PCR amplification, including primer bias (where primers miss certain taxa; or chimera formation (where artificial "hybrid" sequences may be created during PCR amplification.

Bioinformatically extracting phylogenetic or functional marker genes from metagenomic sequence data may offer a less biased method of surveying microbial diversity than metagenetic approaches



So, some of the key points about the advantages of this whole meta genome. So, this is only a part of the whole set of data that can be generated. So, I just took one simple example but there can be many more similar examples or may be more sophisticated and complicated also examples are there. So, some of the key points which are actually; which are to be identified as the major points of advantage are the fact that by examining the functional metabolic attributes of the sequence genome.

It is possible to have a more accurate identification of the microorganism present. Because now we have not only 16s or not only 18s genes but we have the all the taxonomic marker genes. So, based on all the taxonomic marker genes we can actually reach out to all the organisms which are present there and also we have. Now the ability to predict the complete functional profile of the community like the KEG map that I have shown you that is to prediction of the complete functional profile.

Now metagenomic approaches avoid the known biases associated with the PCR amplification including the primary bias where the primer miss certain taxa or chimera formation where artificial hybrid sequences may be created during PCR amplification and they will be thrown out or discarded during the subsequent quality check steps. So, bioinformatically extracting phylogenetic or functional marker genes because here the whole data is with us whole nucleotide and genomic data is with us and it is only bioinformatically the reads are identified and

extracted.

So, they that may offer a less biased method of surveying the microbial diversity than the metagenetic approaches or the target gene amplification approaches.

(Refer Slide Time: 36:16)

One compromise ...sequencing depth

Metagenomics is not targeted
The number of sequences covering any given gene is usually far lower than could be obtained via metagenetic approaches, meaning that far **deeper sequencing** is required for metagenomics to be effective for such applications as biodiversity surveys.

The financial and computational costs associated with acquiring and analyzing these data are often prohibitory.

Nevertheless, metagenomics holds great promise as a versatile technique for studying many aspects of microbial community ecology

However it has one a very severe kind of issue which we must discuss briefly that it is not targeted it is it is going to be for all the organisms in the sample. The number of sequences covering any given gene is usually far lower than could be obtained by meta genetic approach like the PCR based approach. Because here we are not amplifying any specific gene all the genes are available to us.

So, suppose I am looking for a particular carbon fixation gene in my sample but it may so, happen that the carbon fixation gene abundance is naturally very low. So, when I have maybe one million reads and after one million reads are treated I may find out I may found that my carbon fixation gene is not there. So, that possibly means that truly the gene is not there but it could also be possible that if I have like 10 million reads then perhaps I can get the that particular gene detected over there.

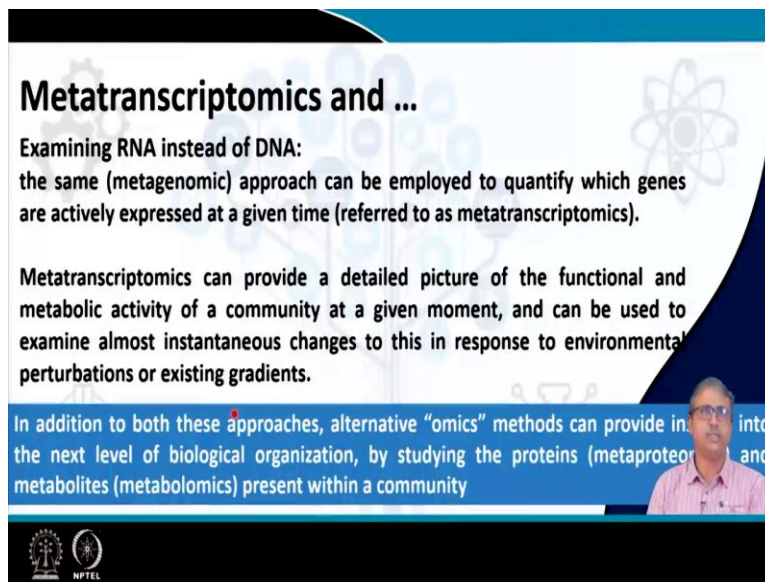
So, meaning that far deeper sequencing is required. So, it is always better to have a deeper sequencing with respect to the whole meta genome sequencing. So, that we do not miss out

anything any important and we are able to actually construct the pathways. Because for each pathway there are certain minimum genes required to complete the pathway. Any pathway would require a set of genes in order to complete the pathway like from glucose to pyruvic acid for example in the glycolytic pathways we need to have all the genes.

If we do not have all the genes in our sample then the pyruvate acid will not be produced. So, the minimum detection is desirable. So, at least one or two copies of the gene must be detected. So, that we are we can say that this pathway is complete. Now increasing the sequencing depth will give us more data. So, that we can quantitatively compare this that what is the quantitative abundance of a particular pathway with respect to another pathway.

Second and minor issue I will say that the financial and computational cost associated with acquiring and analyzing this data for certain research groups and certain countries also these are prohibitory because this is naturally high nevertheless the meta genomics holds great promise as a versatile technique for studying many aspects of microbial community ecology.

(Refer Slide Time: 38:43)



Metatranscriptomics and ...

Examining RNA instead of DNA:
the same (metagenomic) approach can be employed to quantify which genes are actively expressed at a given time (referred to as metatranscriptomics).

Metatranscriptomics can provide a detailed picture of the functional and metabolic activity of a community at a given moment, and can be used to examine almost instantaneous changes to this in response to environmental perturbations or existing gradients.

In addition to both these approaches, alternative "omics" methods can provide insight into the next level of biological organization, by studying the proteins (metaproteomics) and metabolites (metabolomics) present within a community

NPTEL

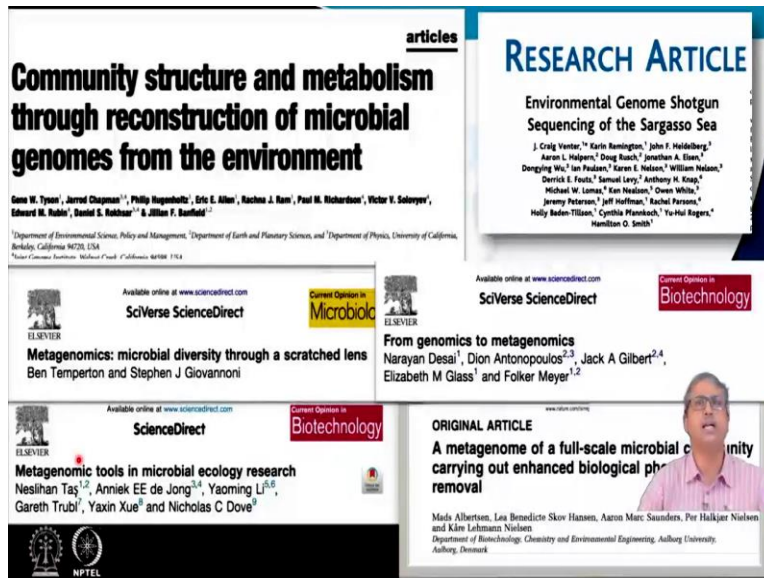
So, meta genomics is not the end of the road following meta genomics we have the meta transcriptomics where we use the RNA so, RNA instead of DNA as I mentioned earlier. So, it is the same approach as we performed with the meta genomic approach that is fragment the pieces or convert the RNA into DNA and go ahead with sequencing but here if we take the RNA then

and apply this technique then we will be able to quantify which genes are actively expressed at a given time and that is called the meta transcriptomics.

And meta transcriptomics can provide a detailed picture of the functional and metabolic activity of the community exactly what is happening in the community at the time of sampling. That is very important because the community might have many pathways, pathways or complete pathways are incomplete many things are there because if we do a DNA based study. But if we do a RNA study will be able to find out very clearly that which pathways are operating under which condition and also the organisms involved in that pathway also.

So, and can be used to examine almost instantaneous changes to this in response to environmental perturbations particularly in the reactors or in the tank where we treat different type of waste water. So, meta transcriptomics is found to be or is actually working very well on with that. Now in addition to both these approaches that i mentioned meta genomics I discussed in detail and meta transcriptomics alternate omics approaches are also there which are called metaproteomics for example considering all the proteins which are present or the metabolomics that is the considering all the metabolites which are present within the community.

(Refer Slide Time: 40:28)



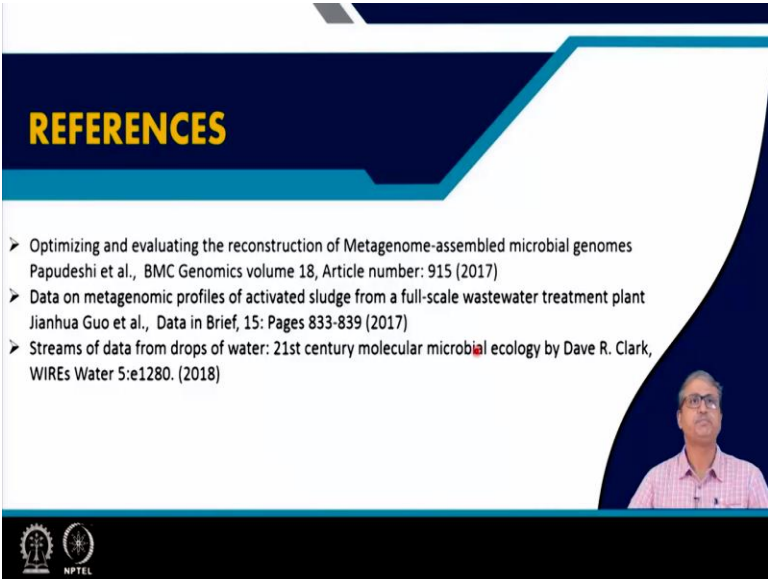
So, here I present some of the very, very important research papers which are which are which are found to be kind of path breaking research I will say. So, like the the first paper in the the left

corner is by Tyson at all. So, this Tyson Nettle and the Venter if you remember the Craig Venter that who actually initiated and is considered to be the stalwarts of this sequencing genome sequencing technology and the process.

So, these two papers the community structure and metabolism and the environmental genome shotgun sequencing of the Sargasso was the initial days paper. So, these papers showed the path that what could be possible and subsequently many, many papers appeared. So, we are not going to have discussion on that but I just want to show you one or two papers which might be important for you if you are interested on understanding how meta genomics tools are being used in microbial ecology research this is a very recent paper 20, 21 paper I believe.

And there is also another paper meta genomics microbial diversity to a scratched lens and discussed certain other points. These two papers by Desai and Albertan published in 2012 are also very useful because they raise certain critical issues and identify the history of development and what exactly meta genomics are able to provide from it from genomics to meta genomics something like that.

(Refer Slide Time: 42:00)



REFERENCES

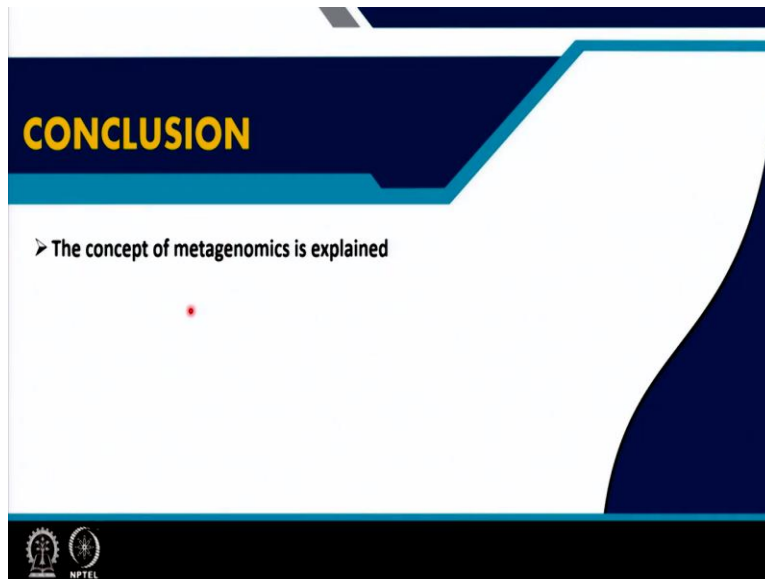
- Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes
Papudeshi et al., BMC Genomics volume 18, Article number: 915 (2017)
- Data on metagenomic profiles of activated sludge from a full-scale wastewater treatment plant
Jianhua Guo et al., Data in Brief, 15: Pages 833-839 (2017)
- Streams of data from drops of water: 21st century molecular microbial ecology by Dave R. Clark,
WIREs Water 5:e1280. (2018)

The slide features a dark blue header with the word 'REFERENCES' in yellow. Below the header is a white area containing a list of three references, each preceded by a blue arrow. In the bottom right corner, there is a small video inset showing a man with glasses and a pink shirt. At the bottom of the slide, there are logos for 'NPTEL' and 'IIT Bombay'.

So, with this these references are given because I found some of these papers are truly very, very useful for you to go through for example the last one the stream of data from drops of water is very comprehensive and also the other papers which will be useful. So, there are numerous paper

and other articles available that can be that can be used.

(Refer Slide Time: 42:20)



So, in conclusion I will say that the concept of meta genomics is briefly explained and it is highlighted that this is the method through which we can actually gain access to the all the organisms taxonomic identity of course but also their functional potential with this I thank you all.