

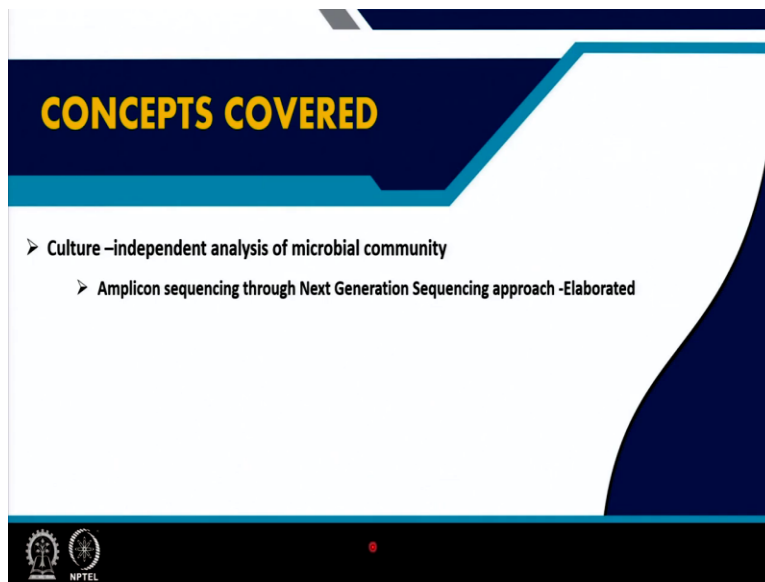
Environmental Biotechnology
Prof. Pinaki Sar
Department of Biotechnology
Indian Institute of Technology, Kharagpur

Lecture – 33

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

Welcome to the next lecture of this environmental biotechnology course and we will discuss in this particular lecture on the methods in microbial ecology with relevance to environmental biotechnology.

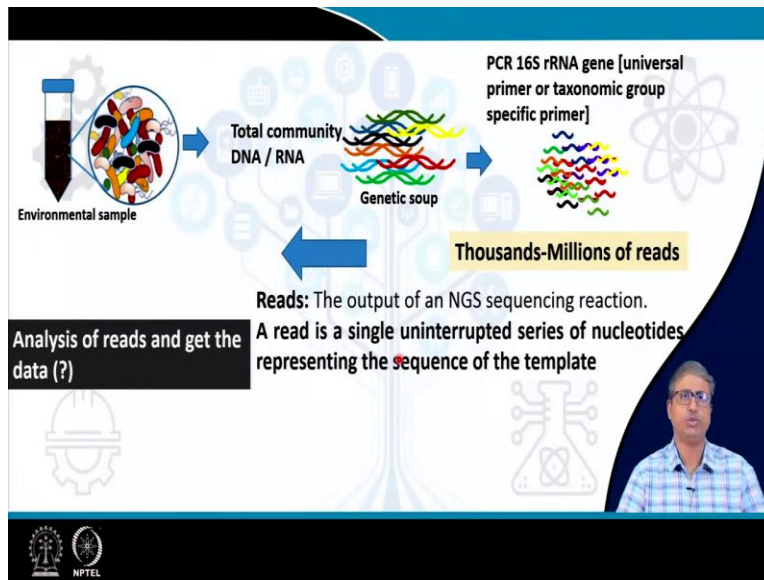
(Refer Slide Time: 00:43)



In particular we will discuss more on the amplicon sequencing based analysis of microbial communities through next generation sequencing approach. We have already discussed some of the aspects of next edition sequencing based approach in our earlier lecture and its comparison with the clone library or other gel electrophoresis based method to differentiate between the different species or organisms present in any environmental sample.

But today we will elaborate uh the scope of this particular technique because in the recent years amplicon sequencing through next generation approach sequencing approach has become one of the major way of characterizing the microbes present in different environmental systems.

(Refer Slide Time: 01:38)



So, as we are discussing already that environmental samples are having most of the time many organisms together and when we extract the total nucleic acid from those samples basically we get this mixture of all the DNA on the genome or the genes or the RNA representing the transcripts or the transcriptomes and this is sometimes referred as the the genetic soup. Now ideally this genetic soup is as you can understand that there are numerous microorganisms numerous bacteria are key and fungi.

So, the genetic soup is going to be a mixture of all these genomes and the genes in particular if we if we want to study the DNA level things. So, this genetic soap as we already discussed that genetic soup needs to be analyzed in terms of the particular target molecules of our interest. So, since we are we are targeting the specific genes which are mostly the taxonomic marker in order to decipher the taxonomy and the composition of the community.

We are taking for example the 16s ribosomal RNA gene but it could be 18s ribosomal RNA gene also if we are planning to characterize the eukaryotic organisms present over there. Or it could be any functional genes any of the functional genes which are which are relevant for the particular process which is of our interest in this particular environment. So, ideally this genetic soup is subjected to a PCR reaction and the PCR reaction is going to have a set of primers and based on the set of primers.

So, if we assume that it is the 16s ribosomal RNA gene that we are targeting. So, we will use a set of primers. So, those primers can be universal primer that means the these primer pair would be able to amplify both bacteria and archaea any bacteria or any archaea present or archaeal 16s ribosomal RNA gene present in the sample or it could be taxonomy group specific primer that means if we want to investigate a particular group of organism that says pseudomonas or bacillus or microbacterium or vibrio depending upon our the research questions or the samples that we are handling.

So, these taxonomic specific primers can also be selected however whatever may be the the primer selection. So, eventually we are going to have a large collection of amplified reads or amplified sequences. Now these are all mixed together if you remember in our earlier lectures when we were talking about clone library or something. So, we were trying to decipher 16s ribosomal RNA genes which are amplified out of the genetic soup using the cloning method or maybe they are segregation by denaturing gradient gel electrophoresis method.

But there are numbers of limitations of those because one of the one of the most important limitation remains that the number of clones that actually can be can be handled can be used and sequenced. So, since we came to know that any environmental sample would have somewhere between several thousands to even millions of species within them. So, it is it is not fair enough to use only a handful of clones or running it in a DGG gel and get a few bands and then and try to try to characterize that that sample in terms of it environmental significance.

So, what we are going to do here we are expecting that from the genetic soup or the mixture of the community DNA or the genomes or the genes when we when we use the 16s RNA gene PCR we will have 1000 to millions of reads. So, what is these reads. So, these reads are basically the output of a next generation sequencing reaction. So NGS sequence sequencing will produce these reads and these are short nucleotide or let us say DNA stretch and they are single uninterrupted series of nucleotides. So, each lead is basically representing a part of the gene ah.

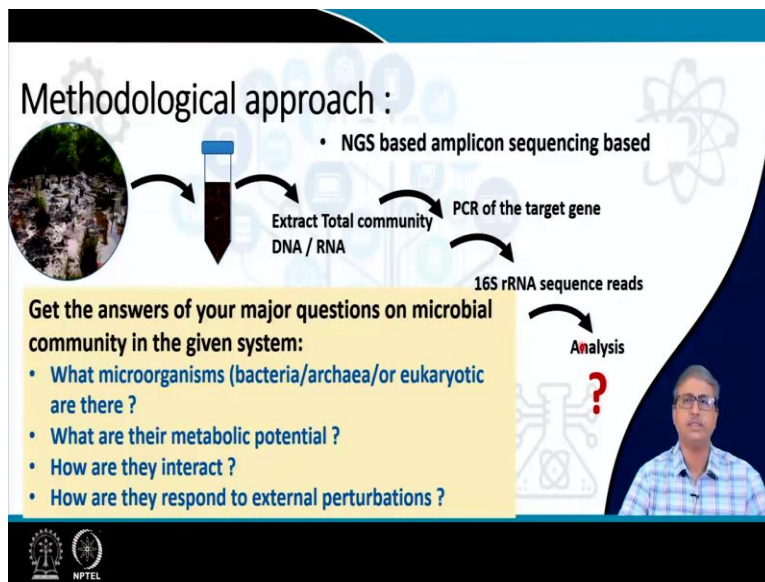
So, if you if we can get may be 1000s or millions of reads that means. So, many 16s ribosomal RNA gene sequences are obtained. So, it has certain certain intrinsic assumption or intrinsic facts

that I will discuss slowly. So, at this moment I want to emphasize that following the amplification of the 16s ribosomal RNA gene or the target gene we will be getting thousands or millions of reads from the sample these are the reads referring to the target gene now what are we going to do with these ribs.

So, we understand that these are reads which are available through the sequencing machine they will appear in our computer screen as discrete read. So, read 1 read 2 read 3 read 4 we can have maybe 10 to the power 5 or maybe more than that number of sequence reads. So, those will be provided by the sequencing machine but we need to understand that how do we analyze and what exactly we want to get from these reads and what is the actually the data that that this can actually provide because the sequence data is not realistically.

So, important unless and until the data is interpreted, so, it is merely a bundle of nucleotide sequences but these sequences are to be analyzed and then the major questions that we are trying to answer using this next generation sequencing based method would be would be addressed.

(Refer Slide Time: 07:54)



Now coming to that point that once we have that large number of reads with us we are posed with a question mark that what is the analysis that we want to do? Now before we come to the analysis procedure the procedures are not the objective of this class. But the objective of this lecture is more towards how do we get the answers of our questions? The questions that we

initially discussed that the questions that are generally asked by any microbial ecologist with respect to an environmental biotechnology system.

That what microorganisms bacteria archea or eukaryotic microorganisms are there what are their metabolic potential? How are they interact? And how are they respond to external perturbations like with respect to change in of the chemical conditions standard change in physical condition and change in other geological conditions within the system. So, for environmental biotechnology answering these questions using the next generation sequencing methods becomes a one of the one of the most important achievement I will say, why?


Because once we deal with so, many sequence reads we are able to achieve a kind of what you call the deep sequencing. So, we are able to decipher the organisms which are otherwise remain unnoticed if we are using clone library or any other methods. It is also very hard to have a kind of a proper picture on how the local environmental conditions are governing them because that is at some point of time is very important in any kind of natural system when we expect microorganism to function and the function should be to our desire that they should function.

So, that we are helped. So, if in order to get that help from the microorganisms we need to understand that how they are able to able to respond to the external environmental factors. So, these are only possible perhaps when we use this next generation sequencing based method and I will give you the examples also that how except actually those are achieved.

(Refer Slide Time: 10:10)

What microorganisms (bacteria) are there in my sample ?

1. How many species are there ? – Number of Operational Taxonomic Units (OTUs)
Species diversity & Richness (Alpha diversity parameter)
2. Relative abundance of the species (expressed as % abundance)
3. What are the taxonomic identities of the species [at different taxonomic levels : Phylum, Class-Family-Genus-Species]
4. Relative abundance of individual taxa (at different taxonomic levels)
5. How does my sample's community composition resemble with others, or samples from the same environment but at different time, condition, etc.



Now coming to the main question that what microorganisms particularly maybe for the ease of discussion we will keep ourselves confined to bacteria that what bacteria are there in my sample. So, that is a very fundamental question the environmental biotechnologist often they ask and we are we are nowhere close to the isolation of few culturable bacteria etcetera because we have crossed those days and we have realized that isolating a few and bacterial strains may not be a very good idea.

In order to understand the totality of the microbes or the microbial strains which are present there. So, this big question there is a bigger question that what microorganisms are there in my sample. So, that is a kind of a very general aspect of the question. Now I can I can decipher or I can actually reset this question into some smaller questions. So, that they are more handy they are more easy particularly when we have a large amount of reads like sequence data that I showed you earlier that millions of reads will be there from a given sample.

So, if we have such huge number of reads actually how this bigger question is going to be answered. So, number one how many species are there? Now here the species refers to the operational taxonomic units. So, these are the most practical definition that we use in the NGS era that kind of based on the sequence similarity or identity we I will come to the definition of operational taxonomic unit.

So, when we define that how many species are there I may have sequence let us say one lakh reads I have obtained one lakh reads. So, out of these one lakh reads how many species are I could detect actually that means how many OTUs are there. So, that could actually help me to understand the species diversity if we want to compare between the 2 samples or if you are monitoring a particular environment with respect to maybe a special or with respect to temporal factors or temporal changes then how this environmental community is getting shifted or getting affected by the changes that will be very useful because the species diversity is very important.

So, species diversity and richness these are 2 common parameters for the alpha diversity analysis of any microbial community. The next question would be what is the relative abundance of the species? Like the individual species suppose I have out of my one lakh reads I have maybe one thousand OTUs 1000 species. So, what is the relative abundance is it that one particular OTU is 10% of the total reads representing the percent wise like 10% or 20% or no it is like just 1% or 0.1% each of these others because there are many OTU's.

So, the contribution of each of the OTUs may be very small but together they represent the entire community. So, the number 3 question or sub question rather what are the taxonomic identities of this species because many OTUs will converge into some taxonomic identity because OTUs are appearing as number like OTU number one or 2 number 2 or 2 number 3 like this. So, that is not a very useful term.

So, we need to get the identification of these OTU. So, when we identify the audios we get to see that many OTUs are actually similar in terms of their taxonomic affiliation at least in the higher taxonomic level like at phylum level or the class level or the family level relatively in the generous level you will find that more number of genera will be there and the species level further more number of species will be there out of the same number of OTUs.

Now relative abundance of these individual taxa are to be determined. So, that means earlier we determine the relative abundance of each of the others. Now since OTU 1, OTU 2 and other all may converge into pseudomonas may be the different species of pseudomonas but all of them are pseudomonas. So, now we have 3 OTUs for example pseudomonas. So, the relative abundance

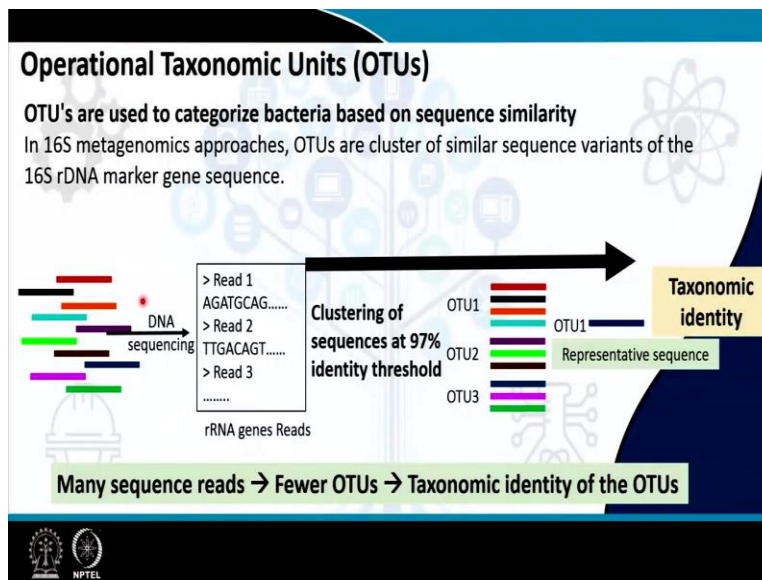
of pseudomonas would be actually the cumulative value of each of the OTUs which represent these pseudomonas.

Similarly all the OTUs that we have we may have 100 OTU's or maybe 500 OUT's. So, the taxonomic identity of each of the OUT's will be determined and then they will be computed. So, that the; taxonomic identity wise abundance can be obtained. So, that can also be at different taxonomic levels. So, starting from the phylum to the class to the family genus and species level next and possibly the last question that I have listed there may be more questions but for simplicity I am restricting only need 5 questions.

How does my samples community composition resemble with others? Many a times it is important to compare the community within the same environment itself or with the similar environment or the environments or the samples which are receiving a different type of exposure to contaminants different type of treatments often aerobic anaerobic. Like in case of the drinking water treatment system we have different kind of reagents or chemicals added to treat the water or the water is passing through different kind of tank system.

So, we need to compare that how the community is actually behaving with respect to the treatments in that case or the conditions or maybe the time. So, different variables may be there. So, ideally so, we may have only one sample but eventually we expect that there will be more samples to compare with and only one when we compare we we get to know many things that that I will come very soon.

(Refer Slide Time: 16:20)



So, the next point that I want to discuss little bit is the OTU concept. So, operational taxonomic units are used to categorize the bacteria based on sequence similarity. In 16s ribosomal RNA gene meta genomics approach that is when we PCR amplify and then analyze through next generation sequencing technique we get a large number of 16s RNA in reads. Now OTUs or operational taxonomic units are cluster of similar sequence variants of the 16s RNA gene or RDNA marker gene sequence at certain threshold identity threshold.

For example you can assume that for a given sample. So there could be some different type of species present there and following DNA sequencing we have obtained a number of reads like read 1 read 2 read 3 and these are the nucleotide sequences which are obtained from the sequencing machine. Now after we get the reads the clustering is done. Now this clustering means the read 1 will be analyzed with respect to read 2 to check that at 97% identity level whether read 1 and read 2 are identical.

It may be identical it may not be identical. So, if they are identical then they will be placed together. So, in this case in OTU one you can see the 4 reads are placed. So, this is just an example that means these 4 reads are the variants of the 16s RNA genes but they have 97% identity among themselves maybe that 3% difference among themselves is that is why they are different otherwise at OTU level 97% OTU level they are forming a same OTU they are clustering together.

So, it is basically determining the sequence identity. So, this in this case the 4 reads are identical at 97%. So, if you consider 100 nucleotides per reads. So, then out of 100 nucleotides only 3 nucleotide differences are there otherwise remaining 97 nucleotides are same exactly the same nucleus right at the same position that is why they are considered that clustered together. So, we have OTU one operational taxonomic unit one.

So, we can assume for time week that this could possibly indicate a particular genera or a particular genus. If we want to delineate species possibly we have to pieces means taxonomic species then we have to increase the threshold to 99% or so. So I will come to that point. So, we gradually try to have this clustering. So, OTU 1 or OTU 2 or OTU 3 and so on and so forth. So, if you have maybe one lakh reads or one million reads you can actually do this clustering.

And we have elegant bioinformatic pipelines developed they are very smart bioinformatic pipeline developed. So, the bioinformatic pipelines will take care about this clustering process and we will taking inputs from you that the whether we want 97% identity threshold or we need 98% or 99% or so, you can you can change the identity threshold in order to have the OUT picking we called OTU peaking a kind of a very, very specific and very particular.

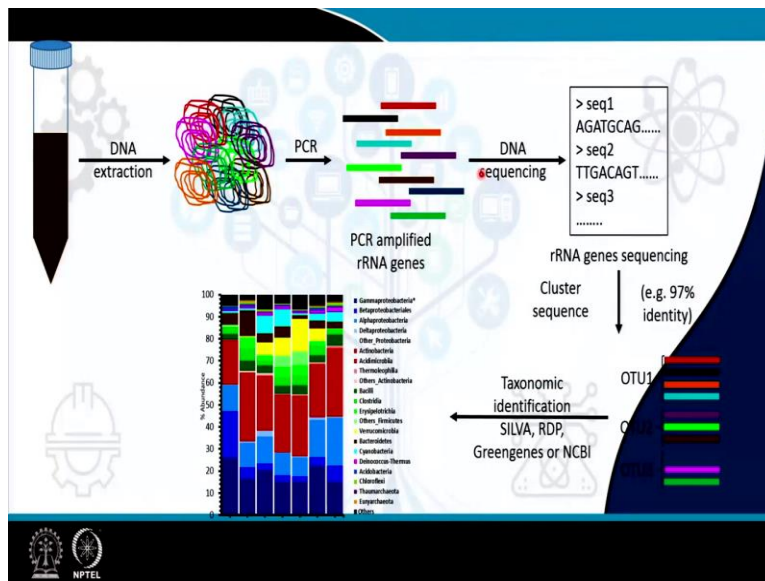
Now out of the reads which are clustered as a particular OTU like in this case as OTU 1 we have 4 reads any one of the reads because we are considering all the reads are similar or identical at 97% threshold. So, any read will be picked randomly this is the algorithm will pick up any read within this particular OTU any read will be picked up that is considered as a representative sequence and will be searched for the taxonomic identity in the NCBI database.

Now this database selection is again up to the in the up to the researcher or the scientists who are working on them. So, there are nice platforms by informatic platforms where all these options will be given that what kind of database you want to use because that there are database called silver database. Database like ribosomal database project or RDP or there are green genes kind of database or NCBI database is always there.

So, there will be options that what type of database you would like to use in order to have a similarity search within the database to identify that OTU belongs to which taxonomic group. So, that means in a summary we are moving like this that many sequence reads are obtained from next generation sequencing method from that relatively fewer number of OTUs are obtained based on the clustering and clustering is done either at the 97% identity or 98 or 99 generally it is 97% then the taxonomic identity of the OTU are obtained.

So, each others will be identified in terms of its taxonomy. So, there are some more points that each of this cluster is intended to represent a taxonomic unit and that is the OTU of a bacterial species or a genus depending on the sequence similarity threshold. In 97% sequence similarity we generally consider a particular OTU as representing a particular genus. We cannot call it a species generally because in order to call it a species we need to have higher threshold like 99% threshold or so.

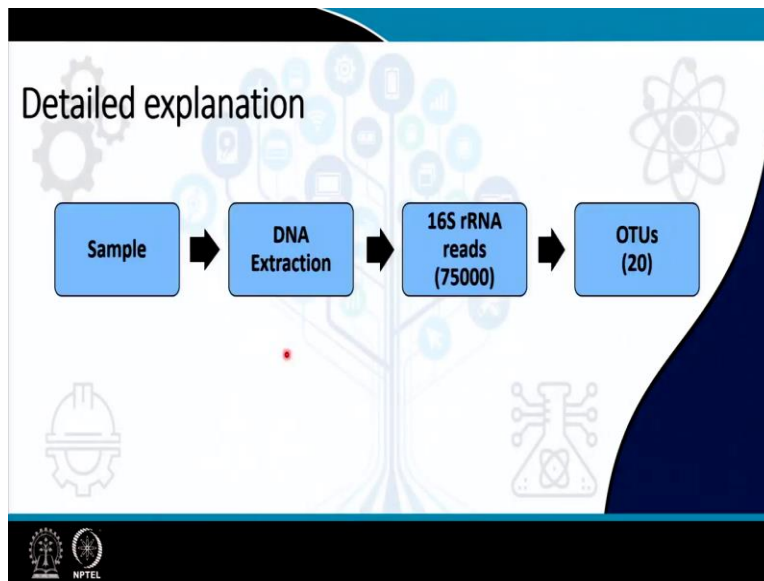
(Refer Slide Time: 21:56)



So, now the essential outcome of this analysis will be that if we have this kind of OTU clustering and each of the OTUs are taken to the database and then identified through different databases like Silver, RDP, Green Genes or NCBI will be able to affiliate each of the others to their respective taxonomy and then those taxonomic information can be compiled and computed and it is it can be graphically represented.

Like here you can see that gamma proteobacteria beta proteobacteria these are at higher level like at class level the information is computed and presented. So, these are the bar graph is representing basically the stack bar are for different stack bars are for different samples. So, we are trying to compare the microbial community between a number of samples. So, we can see that how the community composition remains grossly similar to some extent but there are some very observable differences with respect to certain taxonomy groups.

(Refer Slide Time: 22:55)



Now I will give a simple example that we just we think of a sample any sample environmental sample may be waste water or treatment plant sample just extract the DNA and the that is the genetic soup and go for the NGS sequencing of 16s RNA gene. So, you will have let us assume that 75000 reads are obtained these are these are 16s ribosomal RNA gene rates. Now these reads are subjected to OTU clustering or OTU peaking.

Now it may be possible that the environment is. So, that you may have gone for a huge sequencing like 75000 reads are obtained but they converge or clustered into only a handful of OTU's it is possible.

(Refer Slide Time: 23:37)

Detailed explanation

OTU No.	No. of reads	Abundance (%)	Sequence	Blast match
1	20000	26.66667	TACGTAGGGTCCGAGCGTTAATC	blast match
2	20000	26.66667	TACGTAGGGTCCGAGCGTTAATC	blast match
3	10000	13.33333	TACGAGCACCCGAGTGTGCGG	blast match
4	10000	13.33333	TACGTAGGGTCCGAGCGTTAATC	blast match
5	5000	6.66667	TACGTAGGGGCAAGCGTTGTC	blast match
6	2000	2.66667	TACGTAGGGGCAAGCGTTGTC	blast match
7	2000	2.66667	TACGAGGGTCCGAGCGTTGTC	blast match
8	1000	1.33333	TACGAGGGTCCGAGCGTTACT	blast match
9	1000	1.33333	TACGAGGGGCAAGCGTTAATC	blast match
10	500	0.66667	TACGAGGGGCAAGCGTTGTT	blast match
11	500	0.66667	TACGTAGTCCCGAGCGTTGTC	blast match
12	500	0.66667	TACGTAGGGGCAAGCGTTGTC	blast match
13	500	0.66667	TACGTAGGGGCAAGCGTTGTT	blast match
14	500	0.66667	TACGTAGGGTCCGAGCGTTGTT	blast match
15	500	0.66667	TACGAGGGTCCGAGCGTTGTC	blast match
16	200	0.26667	TACGGGGGCAAGCGTTGTT	blast match
17	200	0.26667	TACGTAGGGGCAAGCGTTGTC	blast match
18	200	0.26667	TACGAGGGTCCGAGCGTTGTT	blast match
19	200	0.26667	GACAGGGGCAAGCGTTGTC	blast match
20	200	0.26667	TACGTAGGGTCCGAGCGTTGTC	blast match
	75000			

So, this is again a kind of a sample data just to explain the point. So, we have the 20 OTU's they are numbered at as OTU number 1 OTU number 20 and the number of reads associated with each other like in the previous example OTU 1 was having only 4 reads. So, but in real situation you may find out the 20000 reads are there in OUT 1 why these 20 OTU's are here these 20000 reads they have clustered at a 97% threshold level.

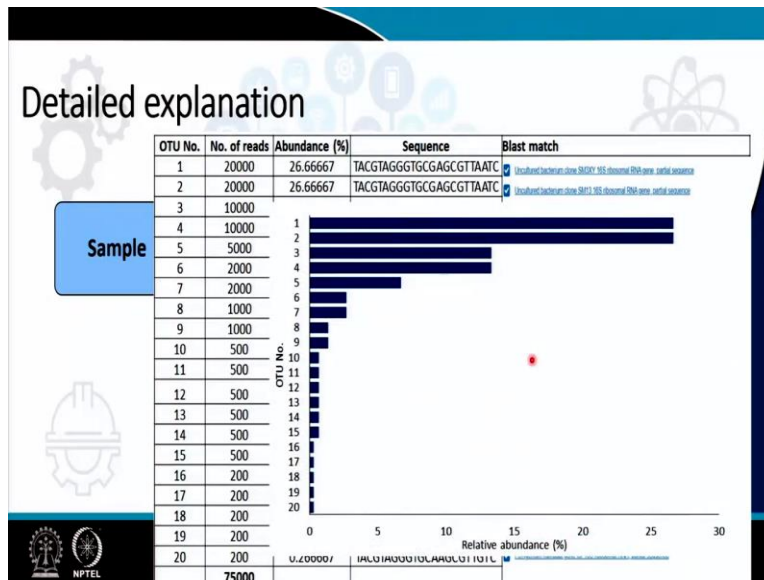
But they are different from the reads which are clustered at as OTU and they are clustered differently with 97% threshold level and from the OTU 3. So, each of the OTU's are actually a different clusters and the reads which clustered as OTU 1 are different from what reads at cluster 2 and different from reads at cluster 3 and so on. So, now if we decrease the threshold like if from the 97% to if we decrease it possibly we will be see that some of the OTUs are actually diffusing to each other.

So, number of OTU's might be reduced if we just decrease the threshold on the other hand if we increase the threshold then we may see the number of OTU will increase because then this will they will disintegrate they will resolve well. Now taking into account that total 75000 reads were there if we just compute that having 20000 read for OTU 1 that means OTU 1 is having a relative abundance of 26%.

So, this OTU which is possibly representing a particular genera or genus with multiple reads

accommodated to it and it has altogether 26% abundance 26.66% abundance. And similarly the relative abundance of all the OTU's are determined and then we can see that the 16s RNA reads for all these OTU's are here and the NCBI blast or the sequence similarity search shows that each of the OTU's are belonging to some organisms. So, these are again just for an example I will come to a real life scenario very soon.

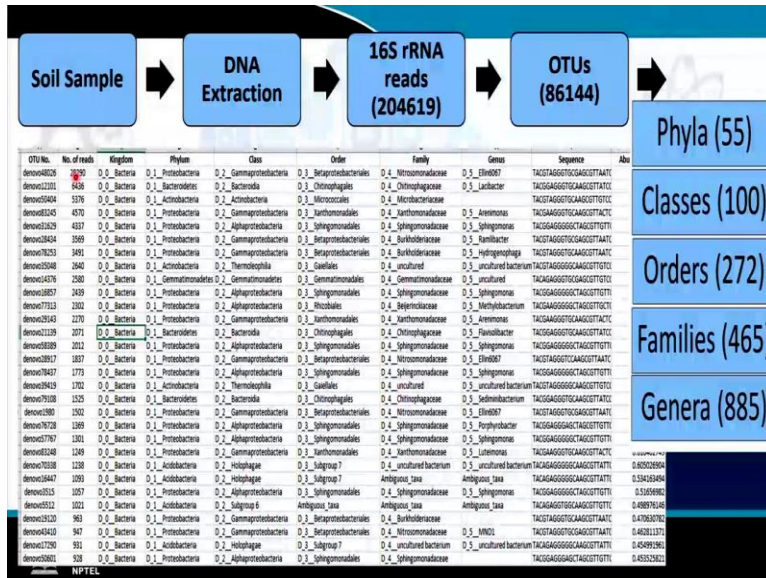
(Refer Slide Time: 25:56)



So, based on this we proceed further to have this OTU distribution plot which is very common this is called ranked abundance plot. In its environmental system this is very useful to understand that what is the OTU's distribution. Many a times what you see that a handful of OTUs are only abundant like in this case we can see that only if 2 4 5 OTU's are most abundant and rest OTU's are just falling down.

They are most of the OTU's are actually having very low abundance that means in a very simple term most of the species are actually having low abundance they are not. So, abundant but there could be many such species present.

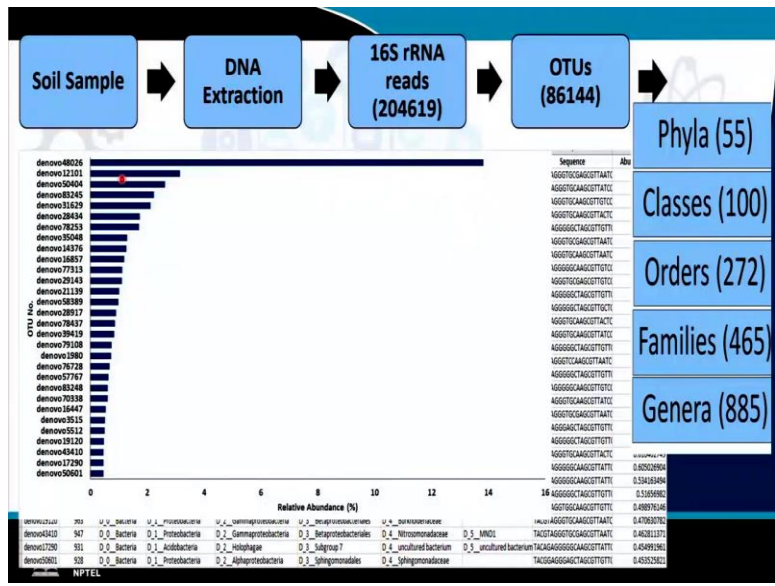
(Refer Slide Time: 26:38)



Now moving forward this is a realistic data where a soil sample which is contaminated soil sample is used to have the same type of analysis and we could get a huge number of reads from the same single soil sample we have got 2,04,619 reads and when we go for OTU peaking or OTU clustering we got 86000 produce a huge number of what use are there that means they are prospective genera or prospective taxa might be.

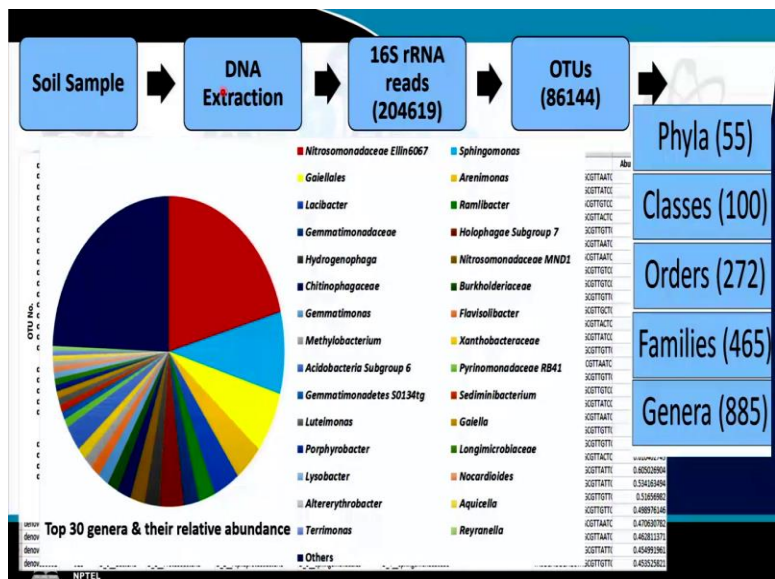
So, we need to look into that. So, when we try to identify them. So, I will show you the data first. So, when we try to look at the OTU data. So, these are the OTU's individual OTU's and if we see that the number of reads associated with these reads are shown here if you look the first de novo is alone having 28000 reads. So, it is a very very abundant OTU compared to the second OTU's which is having 6000 then 5000 and so on and so forth.

(Refer Slide Time: 27:36)



So, if I draw a plot of rank abundance I can surely see that one particular OTU's alone is standing very high it is 14% and rest all OTUs are almost like below 4% or even much less than 2% also. That means it is something like very very characteristic of this particular soil that because of some reason one particular type of organism is very very dominant in this particular sample. So, this is this is actually giving us some kind of ecological or process input that it is it is a very unusual situation in a normal situation this should not happen actually.

(Refer Slide Time: 28:17)



Now if we look at the taxonomy classification of these OTU's we find that when we if we go back quickly I will show you some more interesting things that many of the OTUs are actually being identified up to the genes level but some what use is not identified up to the genes level the

genes level is empty that means this is possibly a kind of insufficient information available in these reads. And the sequence database is unable to assign or it could be a novel taxa.

But it is identified up to the family level that is micro bacteriac but it is not identified at genes table. So, it could be a novel taxa or it could be possible that some more information is required about this particular sequence this short read is not enough. You can also see that for some OTU's like the galilees members it is written uncultured and uncultured bacterium that means again these are previously identified as uncultured and which is quite expected.

And also they are similarly like the previous one they are also possibly novel taxa or they may need some further sequencing effort or the length of the sequencing might be more might be increased. So, what you can you can also see that many OTU's that number of produce are affiliated to Elin 6067 as a general. So, that is the nitrosomonadis generous and followed by this there are also another group which is the Galliellis member.

So, there are couple of Galliellis I can see within this handful of OTUs which are being presented. So, now when we compute all this taxonomic level data like all the nitrosomonadis there are many nitrogenous dc members. So, all this relative abundance of the OTUs belonging to nitrosomonadis are pulled together. Similarly the other families are pulled together we are able to find out how many total families are there how many total genera are there.

And that this is the data that shows that in this particular sample there are total 885 genera are there at general level and there are 465 families are there and there are 55 phylo's are there. Now if I want to see this distribution of this genera that how this genera are distributed. So here we are unable to plot actually all the genera. So, what we could do actually we have plotted only the 30 top genera what is very clear is this red colour followed by the blue and then 2 yellow.

Now what is this red this red is the nitrosomonadis elene 6067 genera or genus. So, that is the most abundant genus followed by the sphingomonas and then the galilees and then the ernie monas type of organisms and also a large number of the members because it is 885 and we could we have only accommodated 30 as I mentioned remaining are accommodated within this. So, a

huge number of genera are there in the sample that is that is very clear.

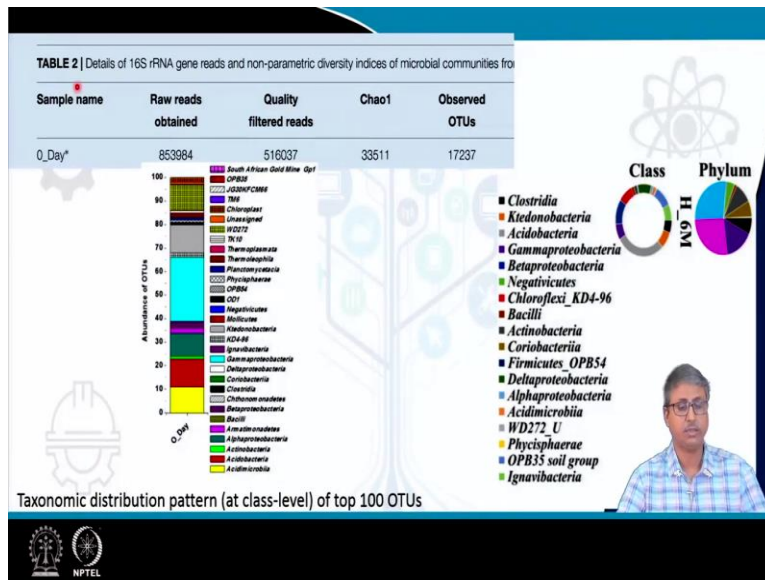
(Refer Slide Time: 31:28)



Now some more practical details that how this next generation sequencing method is actually used in a field scale environmental biotechnology study. In one of our studies we were actually investigating the acid mine drainage site and we are evaluating that how this acid mine drainage environment can be bioremediated. So, this is the acid mine drainage highly acidic water flowing from the mine area.

And this is actually stored in a tank and as you can see that often this water which is very highly acidic water it spills its it spills to the nearby agricultural or the forest area. So, this this you can see the change in colour. So, this is the wet area maybe it may be in a in a 2-3 days ago there was a spillage and lot of lot of acid mine drainage released into the soil area. Now we were trying to analyze that what type of organisms are there and how can we plan a bioremediation process for this.

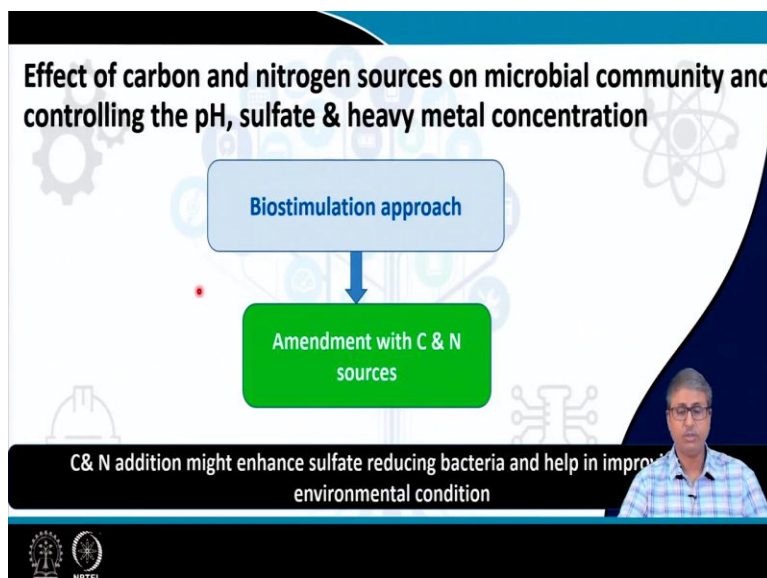
(Refer Slide Time: 32:35)



So, what we did we try to first sequence this sample and as I can I can see that we got 8,53,000 ricks and out of these reads we could observe 75000 OTUs prospective maybe genera we can say and they are distributed into different classes and these are the different classes and when we see the class level data it is very clear that the members of gamma proteobacteria and acid microbia and acidobacteria these are all acidify acidophilic bacteria.

So, they grow very happily in the acidic condition and also the different alpha proteobacteria who are the member of the chemolithotrophic organisms. So, these are the bacteria they are mainly growing in that acid mine drainage impacted soil basically.

(Refer Slide Time: 33:27)

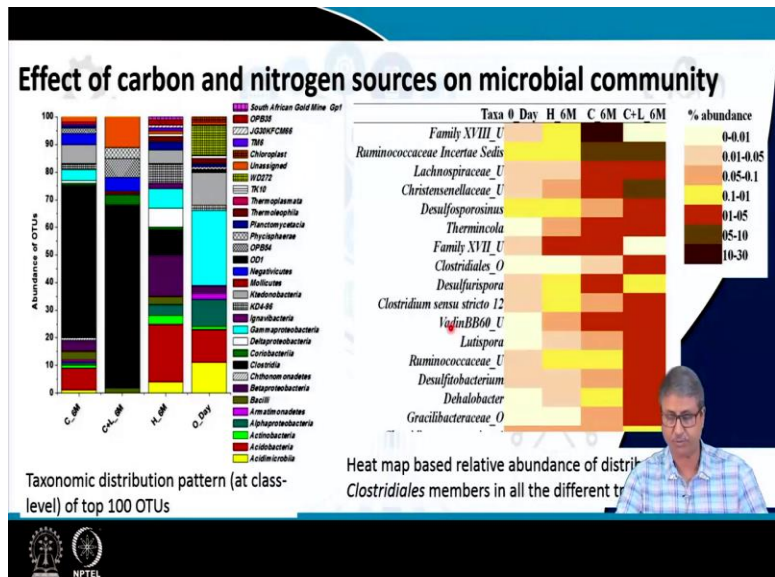


Now since we were trying to have some kind of bioremediation achieved in that. So, the bioremediation is basically achieved in this case by allowing sulphate reduction by native sulphate reducing bacteria. So, we try to adopt the biostimulation approach. So, I will discuss this approach later but right now I will only mention. This is just an approach for bioremediation where some amendments are done no bacteria or marker organisms are added.

It is the natural soil where nitrogen and carbon sources are added. Now it is expected that or the hypothesis is that the carbon and nitrogen sources will allow growth of certain group of bacteria like the heterotrophic bacteria because otherwise this environment is highly autotrophic. So mostly the iron and sulphur oxidizers they grow there. So, you want to grow or allow the growth of the native.

That means the native means the bacteria which are already there in the soil who are capable of heterotrophic growth and utilizing the carbon and nitrogen they should grow and they should do something. So, that the sulphate precipitation occurs by the sulphur reducing process and the pH improves.

(Refer Slide Time: 34:44)



So, so when we try to do that here is the picture that I was referring to you that when you compare I was referring to the fact that next generation sequencing based methods are very useful in comparing the data with the samples same sample with different treatment. So, here we

have the native sample and here we have the sample with the carbon source and the nitrogen source both added and the only carbon source added.

Now you can see the change compared to the native sample the same sample when it is exposed to carbon or carbon and nitrogen source both if this black colour group is found to be very abundant now where from it has it appeared we did not add this bacteria it appeared out of the natural site itself. So, these bacteria if we try to investigate these bacteria might have been there invisible it was.

So, less abundance that is the importance of the relative abundance earlier when we were showing you the plot I was referring to the point that abundance. Now the relative abundance is important because some species might may be very, very less abundant but that may be very useful for our biotechnology purpose environmental biotechnology purpose in particular. So, when we allow a very specific environmental condition in this case we have provided the carbon and nitrogen source we are able to proliferate or support the proliferation of naturally present clostridia bacteria.

These groups of bacteria are capable of sulfate reduction and they can actually improve the pH etcetera. So, if you compare at the family level you can see compared to the zero day sample how the members of different members who are capable of sulphate reduction they have enhanced.

(Refer Slide Time: 36:32)

Major information we obtain from 16S rRNA gene amplicon sequencing

- Species (OTUs) composition (Number & distribution)
- Taxonomic composition
- Relative abundance
- Interaction among different member (of multiple communities)
- Core-community
- Rare-microbiome
- Interrelation with environmental parameters
- Prediction of metabolic function

Now what are then the major information that are expected if we are having a 16s ribosomal RNA gene based NGS method available to us as a tool to characterize our microbial community within an environmental sample. These are the following things that can be readily obtained. Number one the species composition species in the term the OTU level composition there is a number and distribution.

Taxonomic composition that at family level class level generous level whatever level is possible the taxonomic composition, relative abundance of these taxonomic groups like I showed you the example of the cluster DSE member which were originally very less abundant but following a particular treatment those clostridium members where abundance of those members were increased to a very high level.

Next point is the interaction among the different members of the multiple community. So, in order to establish the interaction, so, we need to have multiple samples ah. So, that we can we can use different correlation analysis and based on the correlation among the relative abundance of these taxa maybe at OTU level may be at any taxonomic level we can build the networks from the networks we can understand the interaction pattern.

The next 2 points like the core community and the rare microgram. Core community is the members of the community who are common or conserved or rather they are shared among a

number of samples collected from the same environment. So, suppose we are working on a particular soil environment with contamination and no contamination or different levels of contamination. So, the OTUs which are found to be conserved across the samples are considered to be the members of the core community because they are the conserved member of the community most possibly most stable member of the community.

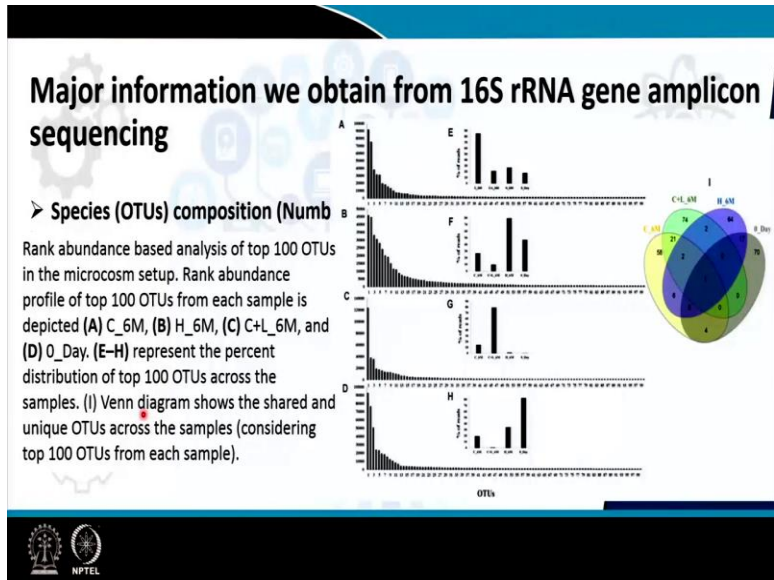
On the other hand the rare microbiome is represented by such OTU's whose abundance is very very low maybe less than 0.1% or so these OTU's are having less abundance but they are more in number if we if we look at the rank abundance plot we will see that these rare OTU's are often found to be very large in number. So, these OTU's are together considered to be representing the rare microbiome.

And in recent time this there is a very high interest in investigating the rare microbiome because otherwise we always try to be more focused on the more abundant populations the more abundant OTU's or more abundant bacteria we rarely talk about the less abundant bacteria but the less abundant bacteria or the species may be more in numbers in terms of individual numbers. So, that is called the rare microbial.

The next point is the interrelation with environmental parameters. So, how the individual community members at OTU level or the taxonomic level they are controlled by the local environmental parameters because then only we are able to identify that how these species or how these organisms would be useful for our environmental application. And the last but not the least is the prediction of the metabolic function.

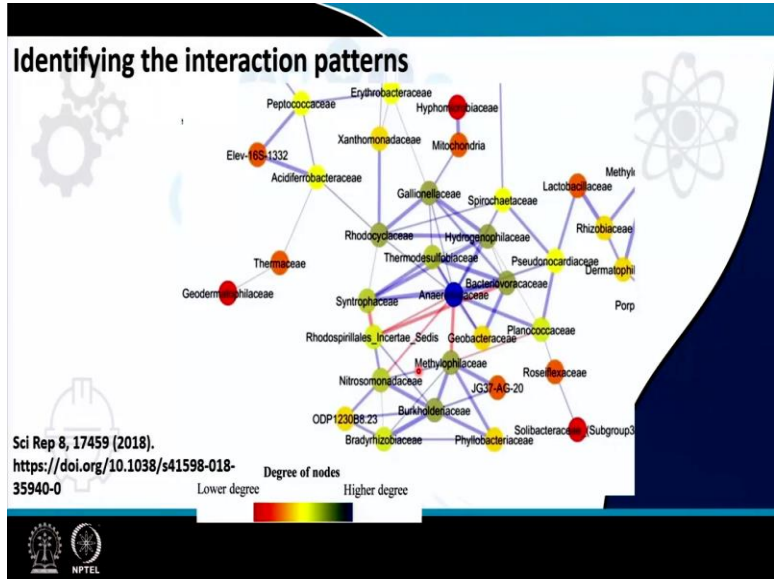
Many a times after knowing the bacterial taxa the name of the genus it is possible to predict their metabolic function. So, there are many methods been developed based on the known genome sequences of the similar organisms. So, those are called predictive analysis of the metabolic function.

(Refer Slide Time: 40:14)



So, I will show you one another data that how this rank abundance based analysis. So, here you can see that the figure D is the native sample and here is the rank abundance. So, these are the OTU's and when we see this in comparison to the treated one like here the A is the carbon or the nitrogen treated and C is the carbon and nitrogen both treated. So, what we see that the OTU's are getting changed. So, this is that means the different species are able to participate in the process as the condition changes.

(Refer Slide Time: 40:56)



Second is the network that I was referring to how the organisms are connected to each other. This is again from another study that we have done earlier where we can see the colour of the nodes are in indicating the more the interactions. So, you can see some anaerobic bacteria which

And we have definite bacterial members responsible for this clustering. Now when we try to see what are the geochemical or environmental parameters that could actually controlling this clustering which is very useful to know that why a particular organism is more abundant in a particular sample what is actually connecting or linking that particular organism to that particular environment this kind of statistical analysis using the microbial community data as well as the environmental data.

So, here you can see the number of variables different chemical and geochemical variable like total organic carbon total carbon or the presence of the potassium or sulphate or calcium how they are actually controlling the the overall community structure in a particular given environment.

(Refer Slide Time: 43:28)

Limitations of 16S rRNA gene amplicon sequencing through NGS

- Choice of hypervariable regions in 16S rRNA gene gives different results
- Taxonomic classification at genus level depending on Database and classifiers
- Limited functional information

Now we are going to have the the limitations of this method although as I mentioned that this method of 16s ribosomal RNA gene amplicon sequencing or 18s in case of eukaryotic organisms have been very is very popular and it is truly very robust method to analyze the community and then also they see for a large number of information. I just pointed out a few like only 5 or 6 major deliverables but one can have other deliverables also.

So, the major limitation that is found is the choice of the hyper variable region of the 16s RNA gene. 16s RNA gene ribosomal RNA gene is basically a long stretch of nucleotide its 1500

nucleotide long. But for new NGS based analysis we generally select a short portion that may be 200 or 300 or 400 nucleotides long. Now which region you are selecting 16s ribosomal RNA gene is having nine variable region or hyper variable region which are used in deciphering the microbial community studies.

Which variable region are you using v 1 or v 2 or v 3 or v 4 these are the name of the variable regions your primary selection will be depending upon the variable regions that you are selecting. So, the result that you are getting might vary depending upon the choice of the hyper variable region. Although in the recent time we have a kind of a consensus on the what region to select. So, we are all may be following the microbial ecologist around the globe they are following a same type of primary selection.

So, there is no ambiguity or no confusion that somebody used a different primer that is why they got a different result something should not; happen like that. The second is the taxonomy classification at genes level depending on database and classifiers I was referring to a number of databases like Silva, NCBI, Green Genes etc. This selection of database is also very important. We need to identify which one is the most robust database for our purpose.

Mostly it is the Silva database which is found to be most robust in the sense it is also very exhaustive in terms of the sequence information stored in that but if somebody is selecting a database like Green Genes then possibly that research group will identify will be able to allocate a lesser number of sequences to genus most of them remain unclassified maybe. And also the third one is the limited functional information because we are we are able to have much information on the taxonomic identity species distribution.

Even we can establish their correlation between themselves or the geochemical or the environmental parameters but we fall sort of functional information functional information means what are the functions of these organisms only some prediction is possible like if you have let us say acid microbial or if you have acidity thy bacillus you might think or you might predict that they are lithotrophic bacteria or if you have species like pseudomonas they are like heterotrophic more versatile type of organisms.

Or if you have a species like genus like geobacter you can think of iron reduction going on in the environment. So, some metabolic functions can be hypothesized or derived based on the prior knowledge. But direct evidence or functional information is not available when you do this 16s because 16s is actually a taxonomic marker we cannot expect functional information from that.

(Refer Slide Time: 47:16)

Limitations of 16S rRNA gene amplicon sequencing through NGS

- ❑ Choice of hypervariable regions in 16S rRNA gene gives different results
- ❑ Taxonomic classification at genus level depending on Database and classifiers
- ❑ Limited functional information

The slide features a background with a stylized tree and various scientific icons. A small video inset in the bottom right corner shows a man in a blue and white checkered shirt speaking. The NPTEL logo is visible in the bottom left corner.

(Refer Slide Time: 47:24)

CONCLUSION

- Details of 16S rRNA gene amplicon sequencing are discussed with real life examples
- Major deliverables and a few limitations are discussed

The slide has a dark blue header with the word 'CONCLUSION' in yellow. The background is white with a dark blue curved shape on the right side. The NPTEL logo is in the bottom left corner.

So, with this I end this lecture and the following references may be suitable for this particular topic. And in the next we have in conclusion that the details of the 16s RNA, RRNA gene amplicon sequencing are discussed with real life examples. And the major deliverables

expectations from the 16s ribosomal RNA gene based amplicon sequencing through NGS along with a few limitations of this process are also highlighted, thank you.