

Applied Environmental Microbiology
Dr. Gargi Singh
Department of Civil Engineering
Indian Institute of Technology, Roorkee

Lecture - 60
Bioinformatics V

Dear students, welcome to our last lecture on bioinformatics in this course Applied Environmental Microbiology. Today we will be continuing the conversation on an actual tool that is available online, anywhere in the globe, to anybody who wants to use it, to analyse the big data that we generate, how to analyse the data we generate from our sequencing and this is very very essential for an applied environmental microbiology engineering student to know.

So, last class we talked about NCBI which has a very important function called blast, where it does a very basic local alignment and it searches the similar most similar sequences or proteins for the entry that we have put, it has various options and I briefly introduced you to them.

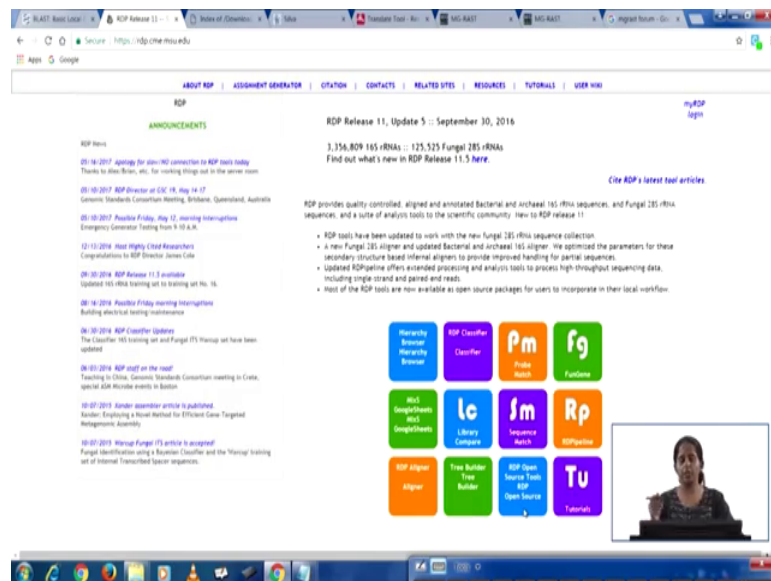
I also should have mentioned to you that sometimes blast p or blast n is better than blast x, sometimes the other way is better than the other. It is very important for you to first of all experiment and try all of them when you generate your sequence and then try to understand if your sequence is so far off a region that is very very established very well accurate like 16 S rRNA, then a blast n would be sufficient do not need to go to blast p also in 16 S rRNA it is the structure of 16 S ribosomal unit that is more importantly.

But if you are looking for a functional gene for example, cellulase which is associated with glycoside which is associated with glycoside hydrolase activity during cellulose degradation. In that case you might consider doing blast p, instead of blast x or blast n or blast x not blast p the blast x, because then it will translate your nucleotides into protein sequences and then blast your protein sequences with protein database.

And that is important because in environment the cellulose degrading enzymes are highly degenerate, which means if there are three different codons that can code for one amino acid. Then probably in some microbes there will be one particular codon use in some another codon and in the some third one. So, if you do nucleotide search we might

survive the gene is not similar because in our database we only have codon one not the other two, but if you translate it into protein and you and do the similarity of the amino acid you will see a lot of similarities. All rightly there are some other options also available to you when it comes to aligning your sequences that you generate from sanger sequencing or the small number of sequences that you generate from anywhere.

(Refer Slide Time: 02:47)

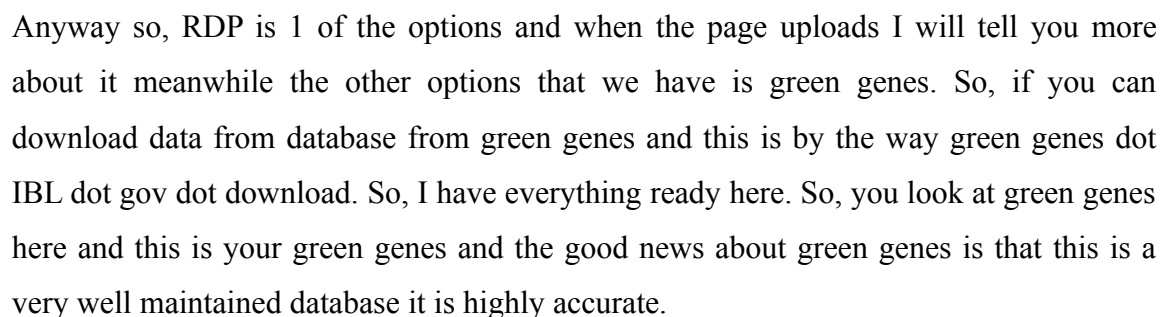


The most popular among them if you are interested in typing and if it were interested in annotating your sequence what and you want to compare and this is when you're comparing the rim ribosomal unit, whether it is 16 S in bacteria or 18 S in eukaryotes and 18 S in archaea remember 16 S in bacteria and 18 S in archaea and fungus in eukaryotes.

So, regardless of what it is we need to you need to the best option that we have is RDP which is ribosomal database project. So, you can just Google ribosomal database project it is maintained by MSU and you can go to sequence match. So, when you go to sequence match make sure that you keep your sequence ready here. So, here we have sequence ready and let this open.

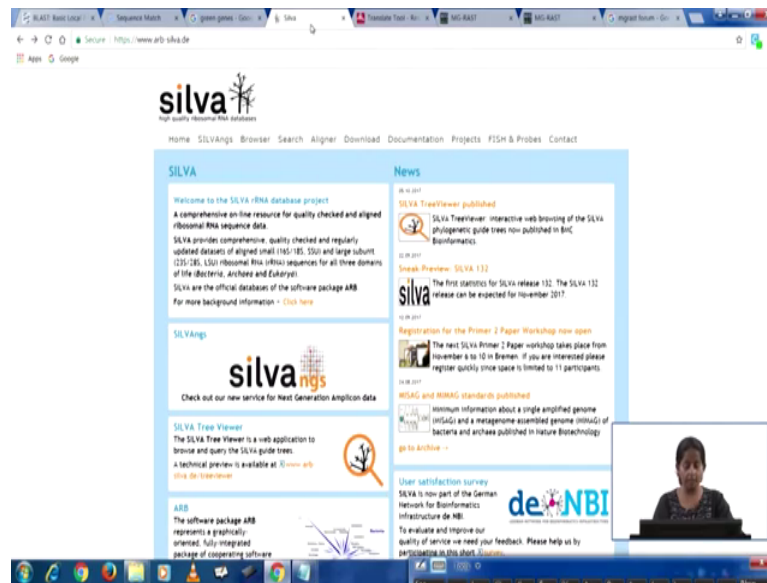
So, once you open this sequence match it will ask you to login in the RDP, but if you I highly recommend that you have an account in RDP if you do not just log in as a guest, but with the guest the problem would be that your data would be not stored. If you log in

(Refer Slide Time: 04:09)



So, if you get a good match with green genes we can be very sure that whatever match we are getting in whatever prediction it is making is probably true. The next option you have a silva is also very well accurate, but not highly limited as green genes.

(Refer Slide Time: 04:45)



So, with green genes the probability of getting a good match is very low, but with silva the probability is pretty nice, you can also download this database and then use bash commands and use different platforms and the information is very well documented in under the documentation centre, but you can find different kinds of tutorials.

The other thing that silva will do is it offers you different different kinds of tools and you can go through them very quickly. It will allow you to do the allow you to download the database it also has an silva ngs, where you can use your high-throughput sequencing data. And then with your high-throughput sequencing data you can use silva NGS to make sense out of it we also have silva tree viewer.

So, many at times many programs will allow you to make a tree, but you cannot view it. So, remember the tree here is your fellow genetic tree how similar the sequences or samples are to each other. So, they will give you a symmetric, which will have all the data you need to actually draw a tree a dendrogram cluster dendrogram, but it will not make one. So, if you have silva tree viewer you can input your file. And then it will make a tree for you.

(Refer Slide Time: 05:52)

The screenshot shows the SILVA website homepage. The main content area includes a 'SILVA 128 release' announcement, a 'User satisfaction survey' by deNBI, and a 'SILVA SSU / LSU 128 - full release' table. The table lists various parameters like 'Minimal length', 'Quality filtering', 'Guide Tree', 'Release date', and 'Aligned rRNA sequences' for different database versions. A 'SILVA on Twitter!' section is also visible. A video feed of a presenter is in the bottom right corner.

	SSU Parc	SSU Ref	SSU Ref	LSU Parc	LSU Ref
Minimal length	300	1200-1900	1200-1900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	no	yes	no	yes
Release date	28.09.16	28.09.16	28.09.16	28.09.16	28.09.16
Aligned rRNA sequences	5,616,941	1,922,213	645,151	735,238	154,297

Then we have ARB, which is a big project and then here they have UniEuk, which is universal taxonomic framework for eukaryotes and then and another very important part of silva is the sina online.

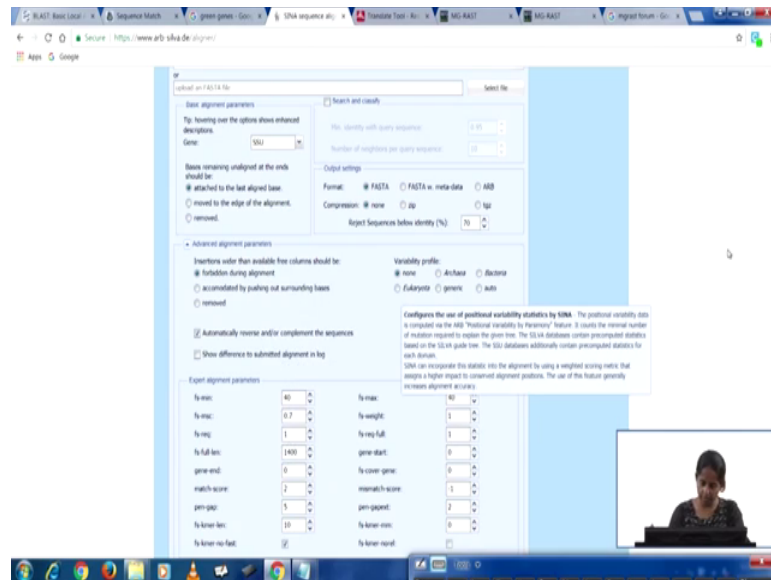
So, sina online is very much like NCBI. So, if you have few sequences like 1 2 3 4 5 6 sequences not a lot you can do sina alignment. So, you paste a fasta file here.

(Refer Slide Time: 06:09)

The screenshot shows the SILVA Incremental Aligner (SINA) web interface. The page features a 'FASTA file' input field with a sample sequence. Below it are 'Basic alignment parameters' and 'Advanced alignment parameters' sections. The 'Basic alignment parameters' section includes a tip, a name field, a minimum identity threshold, a number of neighbors per query sequence, output settings, and an alignment threshold. The 'Advanced alignment parameters' section includes a job name, an aligner selection, and a 'Run aligner' button. A 'SINA Incremental Aligner (SINA) Summary' section is at the bottom.

So, we have pasted a fasda file here right perfect and you can also upload a file if you have different if you have a file properly maintained. And then you can choose different options this is SSU gene yes we want it attached with last align base or you want it attached to the ok.

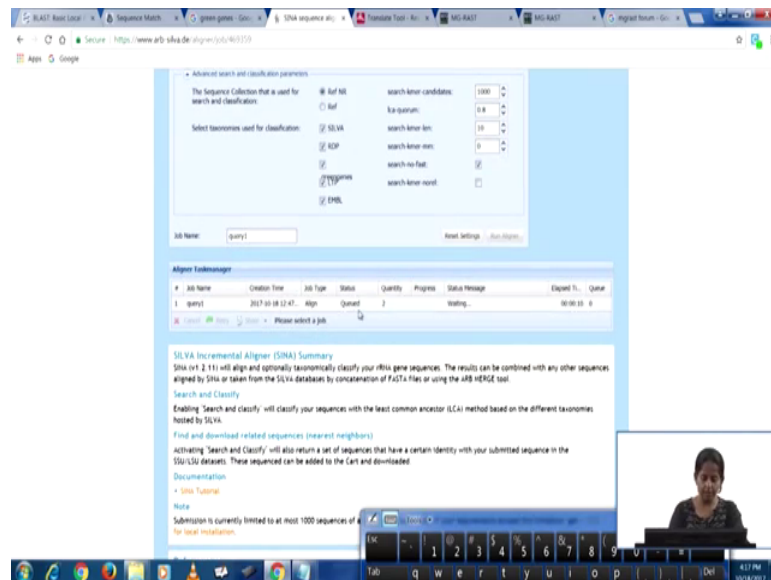
(Refer Slide Time: 06:43)



Last line just leave it at last line base and then, you can also look at the advanced alignment parameters, we are expecting it to be bacterial so, um, but anyway we can just write now. All righty and you can give it a name if you want query one very good.

And, then these are different parameters you can change them as per your requirement and if you click search and classify what it will do is it will not only search the similarities to all other silva database entries, but it will also classify your sequence for you, it will tell you whether it is coming proto bacteria or not, but remember the first one according to NCBI was gamma proto bacteria and second will (Refer Time: 07:34), let us look at what this is for let us run aligner already. So, our query one is under Q.

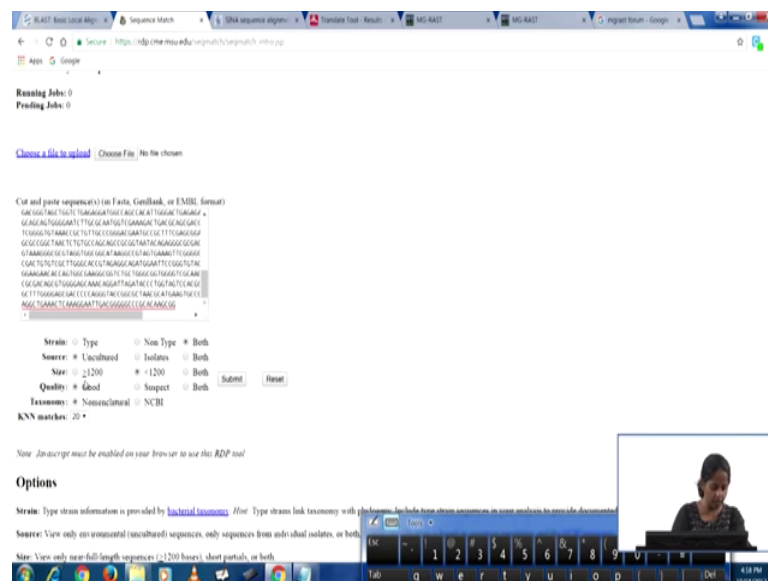
(Refer Slide Time: 07:43)



So, when this query one will be done we can actually open it and we can take a look we can show the result we can download the file and we can add our neighbours to the card anyway. So, sina is a very good tool available from silva for amplicon sequencing.

And this is green genes I highly recommend both of them and let us see what an RDP ok. So, RDP this is not opening properly.

(Refer Slide Time: 08:08)



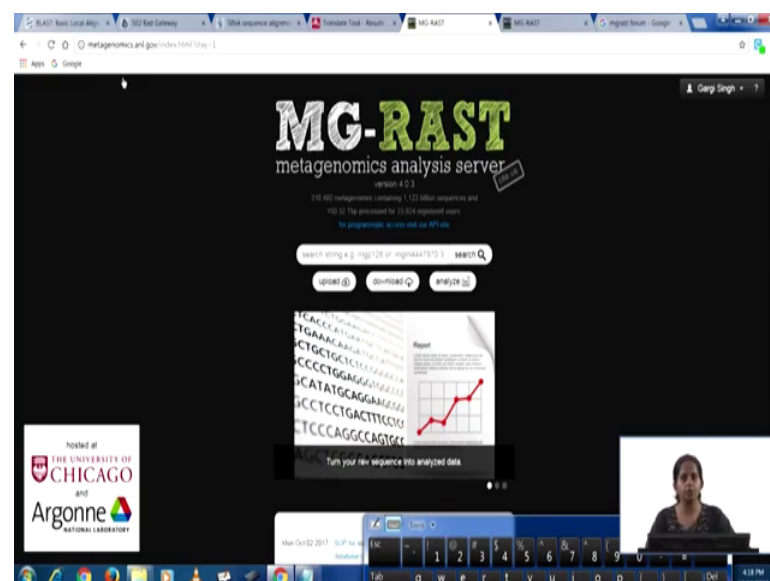
But you can cut and copy paste your sequences here, you can tell them what kind it is it is uncultured the size is less than 1200 this pair quality is good we are interested in nomenclature. And then you can have different kinds of options, and then you can submit it and then it should also give you data ok.

(Refer Slide Time: 08:30)



And then this is the translation tool by the way just to remind you expasy is the most popular one.

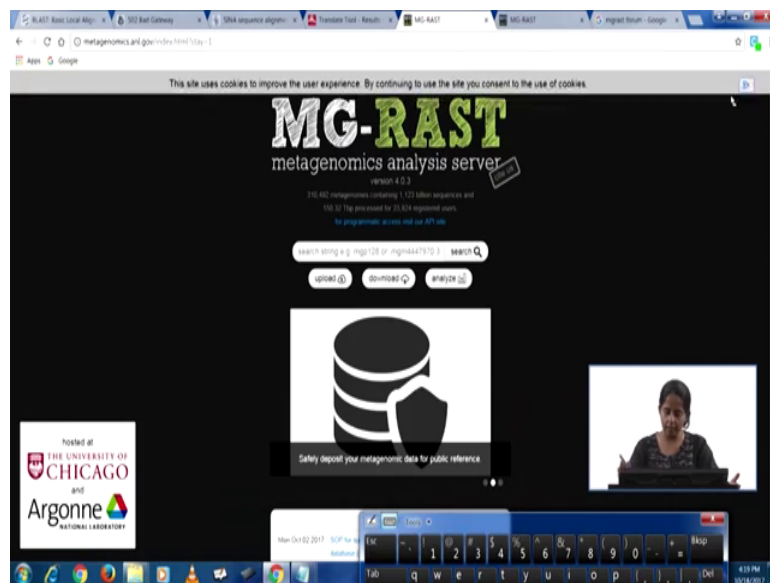
(Refer Slide Time: 08:35)



Now, let us come to high throughput sequencing. So, in high-throughput sequencing what happens is we generate a lot of data tremendous amount of data. And so, we have thousands or millions of sequences and there is no way that we can use an online tool like NCBI or like sina or like RDP to classify our sequences. Another thing is with high throughput amplicon sequencing, when we are talking about amplicon sequencing not the whole genome sequencing.

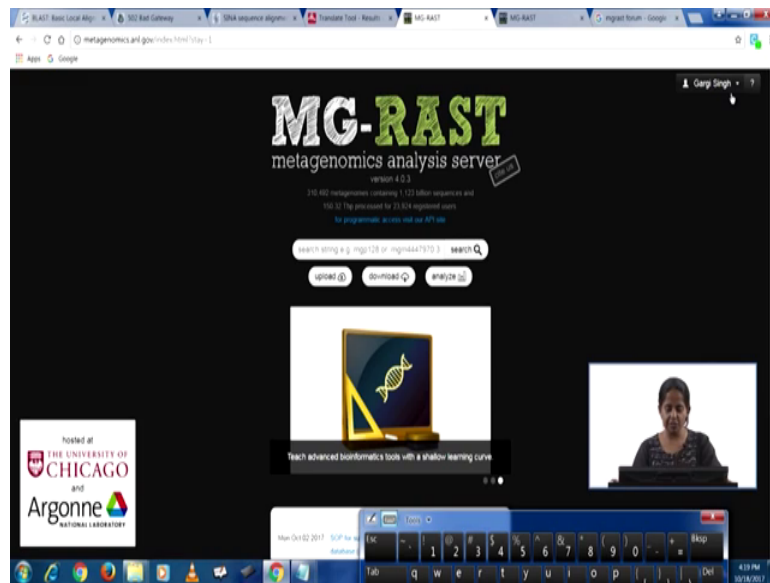
So, if you are not clear about what is the difference between to I highly recommend go back to previous lectures, figure it out and then watch this lecture ahead all rightly. So, if you are watching it now you you know the difference between high throughput amplicon sequencing and whole genome sequencing. So, in high throughput amplicon sequence in this process we had a lot of errors, we had a lot of chimeras that may come up.

(Refer Slide Time: 09:33)



So, we require a dedicated tool to analyse it for us and then give us the data a meaningful data MG RAST is supply online is an online platform, which is very popularly used it is almost the industry standard right.

(Refer Slide Time: 09:39)



Now, and it is a Meta genomics analysis server that is what they call themselves. And what you can do is you can upload your data? You can allow it to analyse clean your data remove the chimera, remove the errors, then you can download the results and then you can analyse it here too. So, one of the first things you need to do is you need to login and in order to login you need to give them requests.

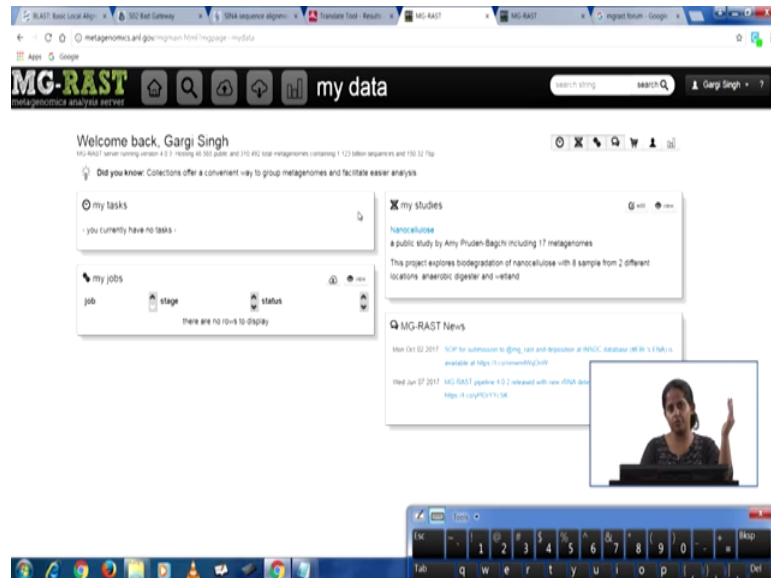
And, then they take few days to get back to you see please do not leave anything for last minute because MG RAST remember we're dealing with high throughput amplicon sequences. So, we have lot of sequences, lot of data, it will take anywhere from few days to weeks for the MG RAST to finish it is processing of your data.

And MG RAST gives you much multiple options you can choose to make your data public upon completion of and this is by MG RAST in that case they will give you priority, because they promote transparency and free sharing of data or you can say that I want to make my data public after 6 months or after this much time. Then you will be next in order on priority list and then you can say I do not want to make my data public at all until I choose other right, then you will be under least priority.

So, you can choose what level of privacy you want for your data and then your priority and the queue will also be affected all right let us look at some example. So, this is mine MG RAST and what I want to do is I want to see what is going on with my data?

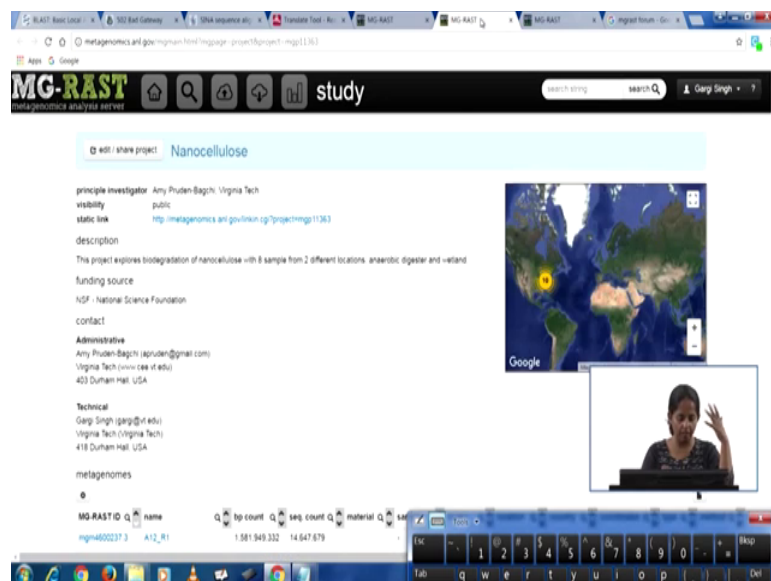
So, I can click on my data and it will bring my data, but let me warn you MG RAST tends to be slow, because it is dealing with very heavy very heavy data and a lot of data.

(Refer Slide Time: 11:10)



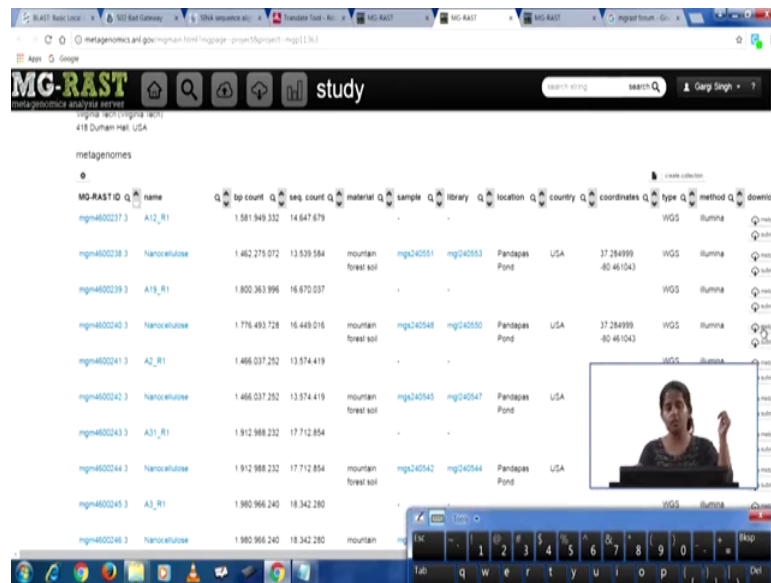
So, the server tends to be a little slow. So, please do not look if you're going to if you have amplicon sequencing data do not leave your analysis for last minute. So, you can click on the Nano cellulose here.

(Refer Slide Time: 11:31)



And this is my project by the way where I have clicked Nanocellulose project.

(Refer Slide Time: 11:33)



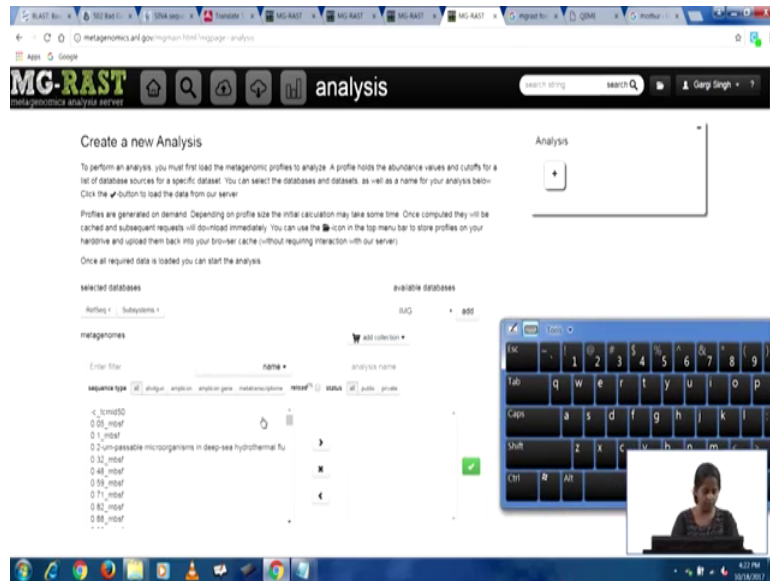
MG-RAST ID	name	bp count	seq count	material	sample	library	location	country	coordinates	type	method	download
mgm4000237.3	A12_R1	1,581,949,332	14,647,679	-	-	-	-	-	-	WGS	illumina	download
mgm4000238.3	nanocellulose	1,462,275,072	13,539,584	mountain forest soil	mg240551	mg240553	Pandapas Pond	USA	37 284999 -80 461043	WGS	illumina	download
mgm4000239.3	A19_R1	1,800,363,996	16,670,037	-	-	-	-	-	-	WGS	illumina	download
mgm4000240.3	nanocellulose	1,776,493,728	16,449,016	mountain forest soil	mg240548	mg240550	Pandapas Pond	USA	37 284999 -80 461043	WGS	illumina	download
mgm4000241.3	A2_R1	1,466,037,252	13,574,419	-	-	-	-	-	-	WGS	illumina	download
mgm4000242.3	nanocellulose	1,466,037,252	13,574,419	mountain forest soil	mg240545	mg240547	Pandapas Pond	USA	-	WGS	illumina	download
mgm4000243.3	A31_R1	1,912,988,232	17,712,854	-	-	-	-	-	-	WGS	illumina	download
mgm4000244.3	nanocellulose	1,912,988,232	17,712,854	mountain forest soil	mg240542	mg240544	Pandapas Pond	USA	-	WGS	illumina	download
mgm4000245.3	A3_R1	1,980,966,240	18,342,280	-	-	-	-	-	-	WGS	illumina	download
mgm4000246.3	nanocellulose	1,980,966,240	18,342,280	mountain forest soil	-	-	-	-	-	WGS	illumina	download

And, then we will wait for it to come up and then these are the sequences that had submitted note these sequences are not high throughput amplicon sequencing sequences these sequences are whole genome sequences. So, MG RAST is for whole genome and for high throughput amplicon sequencing the two most popular are mothur and qiime it will be very briefly talking about mothur in qiime all righty.

So, what I can do is I can pick 1 of these sequences for example, let us say I am interested in A12 R1 and this is a sequence that I had submitted and MG RAST took a week or so, to analyse the sequence and then let us see what kind of output data MG RAST will give us.

So, this is going to take a bit too open up meanwhile what I will do is I will open up for qiime and mothur not lately what another thing that I can do is.

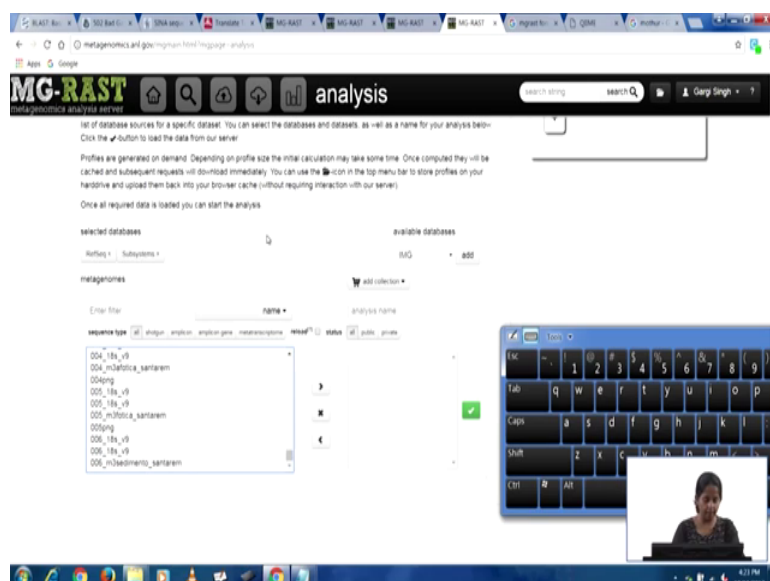
(Refer Slide Time: 12:32)



I can look at I can now here I have different options than MG RAST this is home takes me back to the original page ok. And then this is searched where I can search different thing this is upload data if I want to upload my data does it download my data freely available data and this is analysis.

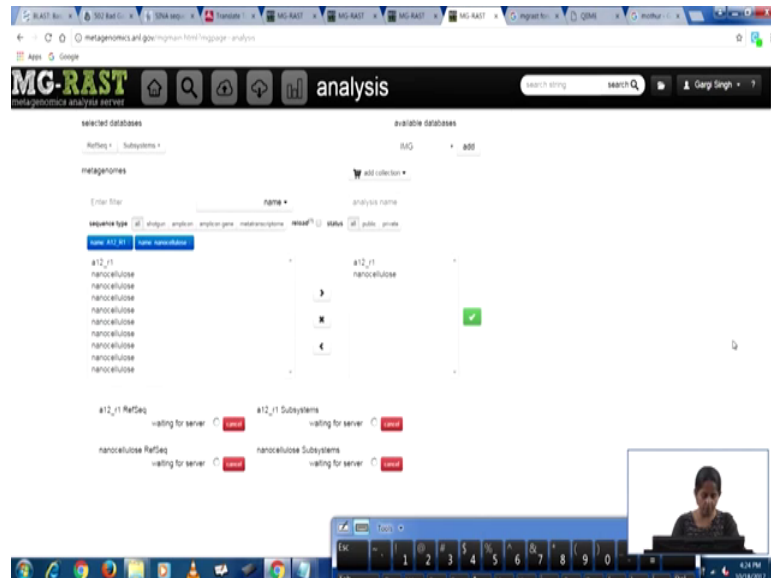
So, I have to do an analysis to click on analysis it will give me a long list of what you what I want to analyse and I can choose. So, let us say I want to analyse my own data.

(Refer Slide Time: 13:00)



Let us say I to analyse A 1 2 dot R 1. So, I am going to type here underscore let us see if this is there all righty.

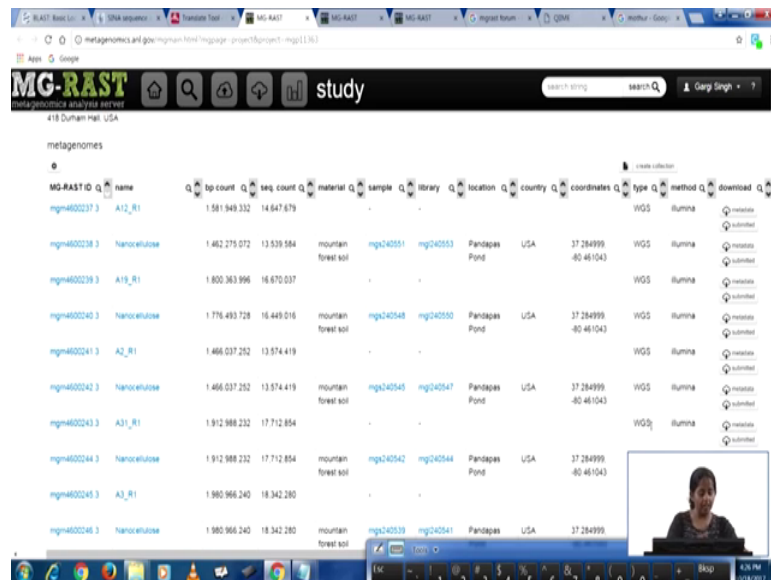
(Refer Slide Time: 13:50)



So, I have 1 file called A 1 R 2 I can select this file the other file let us write I want to look for nanocellulose, because I know some of my files are generically referred to as nanocellulose all righty. So, I will take this one and I have accepted two files. So, now, what I can ask the md RAST is to do yes I want to compare these two data so, these two samples.

Now, if this is going to take some time. So, let us move on here no progress yet. So, let us see what is up with our sina again this is going to take time. So, let us say I want to look at my this particular file. So, I will open this file and here look it is actually even telling me how much how much of data it has processed. So, far how much it has not processed one thing you want to note as you are waiting for MG RAST to open up the data, because as I mentioned that data is huge.

(Refer Slide Time: 14:58)



The screenshot shows the MG-RAST web interface. At the top, there's a navigation bar with the MG-RAST logo and a search bar. Below the navigation bar, there's a table of metagenomic samples. The table has columns for MG-RAST ID, name, bp count, seq count, material, sample, library, location, country, coordinates, type, method, and download. The data is filtered to show samples from the 'metagenomes' category. A video feed of a person is visible in the bottom right corner of the browser window.

MG-RAST ID	name	bp count	seq count	material	sample	library	location	country	coordinates	type	method	download
mgm600237.3	A12_R1	1,581,949,332	14,647,679	-	-	-	-	-	-	WGS	Illumina	download
mgm600238.3	nanocellulose	1,482,275,072	13,539,584	mountain forest soil	mgc240551	mgc240553	Pandapas Pond	USA	37 284999 -80 481043	WGS	Illumina	download
mgm600239.3	A19_R1	1,800,363,996	16,670,037	-	-	-	-	-	-	WGS	Illumina	download
mgm600240.3	nanocellulose	1,776,493,728	16,449,016	mountain forest soil	mgc240548	mgc240550	Pandapas Pond	USA	37 284999 -80 481043	WGS	Illumina	download
mgm600241.3	A2_R1	1,486,037,252	13,574,419	-	-	-	-	-	-	WGS	Illumina	download
mgm600242.3	nanocellulose	1,486,037,252	13,574,419	mountain forest soil	mgc240545	mgc240547	Pandapas Pond	USA	37 284999 -80 481043	WGS	Illumina	download
mgm600243.3	A3_R1	1,912,988,232	17,712,854	-	-	-	-	-	-	WGS	Illumina	download
mgm600244.3	nanocellulose	1,912,988,232	17,712,854	mountain forest soil	mgc240542	mgc240544	Pandapas Pond	USA	37 284999 -80 481043	WGS	Illumina	download
mgm600245.3	A3_R1	1,980,966,240	18,342,280	-	-	-	-	-	-	WGS	Illumina	download
mgm600246.3	nanocellulose	1,980,966,240	18,342,280	mountain forest soil	mgc240539	mgc240541	Pandapas Pond	USA	37 284999 -80 481043	WGS	Illumina	download

So, it takes time is the kind of the kind of information they require from you, when you are submitting your data they need to know what is what name you want to use and then; obviously, they can calculate the base pair count of sequence count by themselves then you need to inform them what kind of sample it is, where did you collect the sample from which country, what are the coordinates, what kind of sequencing you did and what is the method of sequencing?.

So, this metadata they required because as I mentioned they promote free sharing of data. So, if I just have the file here without knowing where it was collected from when it was collected what kind of sample it is and what kind of analyses was done, what kind of sequencing was done, then I cannot use the data for my own analysis. So, this metadata is very very important for us without metadata it will not accept your sample you need to submit metadata.

(Refer Slide Time: 15:47)

The screenshot shows the MG-RAST (Metagenomics Rapid Analysis Server) interface. The main header displays the MG-RAST logo and the text "shotgun metagenome". Below the header, there is a search bar and a user profile for "Gargi Singh". The main content area is titled "A12_R1" and includes a description: "This shotgun metagenome is part of the study 'Nanocellulose' by Amy Pruden-Bagchi, Virginia Tech". A table lists various identifiers: ID (d0911912008/RA0143030303213172633), ENA Project ID, Static Link, Sample, NCBI Project ID, GOLD ID, ENA Library ID, PubMed ID, and Library. A "Table of Contents" sidebar on the right lists navigation options: Home, Analysis Statistics, QSC Mx3 Info, DRISSE, K-mer Profile, Nucleotide Histogram, Source Hts Distribution, Functional Hts, Taxonomic Hts, Rank Abundance, Rarefaction Curve, Alpha Diversity, Sequence Length Histogram, Sequence GC Distribution, Sample Data, and Download. The main text area contains detailed information about the data set, including the upload date (2014-11-09), total sequences (14,647,679), and a breakdown of sequences that failed QC (1,569,593, 10.72%), were unknown (1,207,072, 8.24%), or were predicted features (12,071,014, 82.41%).

So, for now I looked up the sample A 1 to R 1 and I wanted to know what are the qualities of the sample? So, I clicked on it took a bit for the third page to upload, but this is the kind of information it will give you post analysis of your sample.

So, it knows that I chose privacy I chose that I do not to share this information, because I was not sure when I am going to publish this and then it tells.

(Refer Slide Time: 16:15)

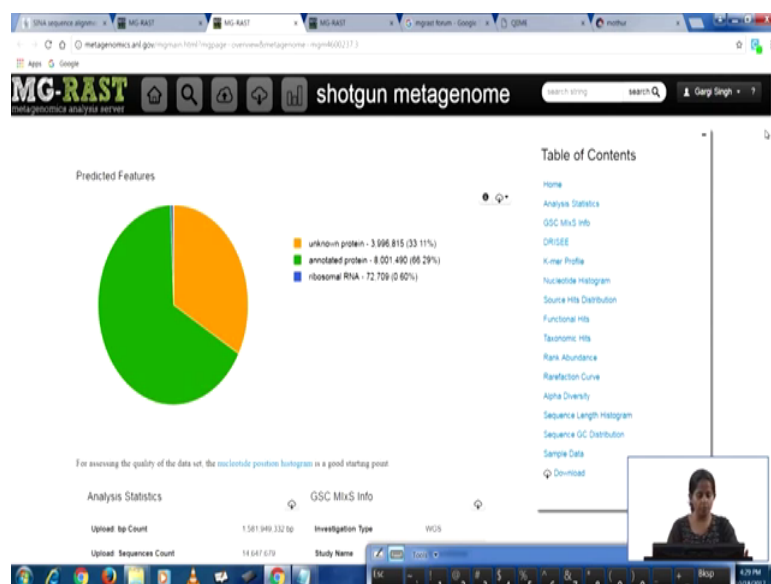
This screenshot shows the same MG-RAST interface as the previous one, but with a "Sequence Breakdown" pie chart prominently displayed. The chart is divided into three segments: a large purple segment for "predicted feature" (12,071,014, 82.41%), a smaller orange segment for "unknown" (1,207,072, 8.24%), and a small grey segment for "failed QC" (1,569,593, 10.72%). The "Table of Contents" sidebar remains on the right, and the main text area contains the same detailed information about the data set as in the previous screenshot.

Who was the lead pi of this research and then look here it will give more into a generic information about the sample the dataset was uploaded on this time at this date it has these many sequences these many base pairs the after going to pair 8 base pair. And then if you give us a basic summary of how many what percentage of sequences did not pass the quality control pipeline, how many had D replication and the ones that passed?.

How many had RNA ribosomal RNA sequences? How many had put? How many had sequences that are that we know are associated with proteins with predicted functions and then how many had proteins predictor proteins that we do not know the function of and this basic summary it gives and then it draws some graphs that are very helpful same information, but written in graph for. So, it will give me a sequence breakdown it will tell me there what percentage 10.72 percent failed quality analysis quality control.

6.88 percent is unknown this is important these sequences could be chimera, because they are not known they are not present in the database. It is also possible that these are some different kinds of gen genes that we have no idea of yet and these are the ones at 2.4 percent that we predicted, that we know we have information in database off.

(Refer Slide Time: 17:29)

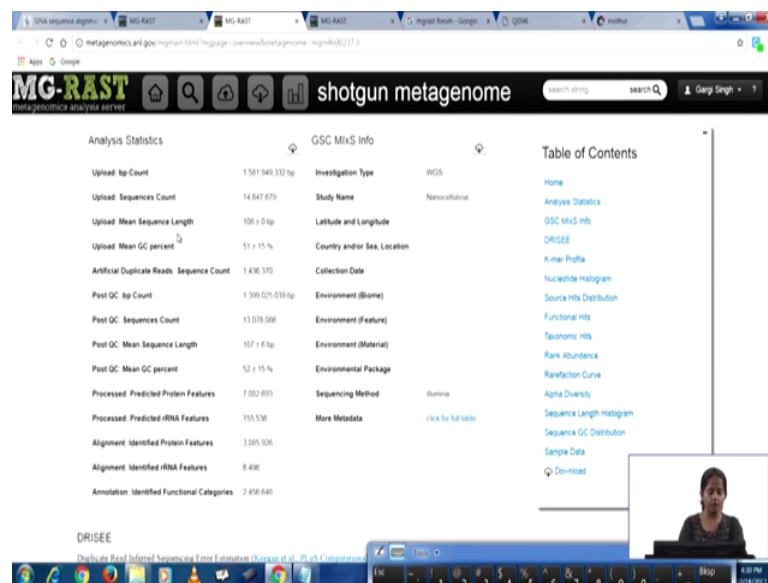


The second one tells about predicted functions. So, majority of the proteins predicted proteins had ever annotated. So, we know what their function is, but 33 percent one third of times we do not know what the function is a very tiny portion of it is ribosome. So, if you want to find out what kind of microbes are present in your sample going for whole

sample meta genomics may not be very good idea, because the you will be sequencing lot other more proteins a lot other more DNA than just one that stands for 16 S RRNA gene.

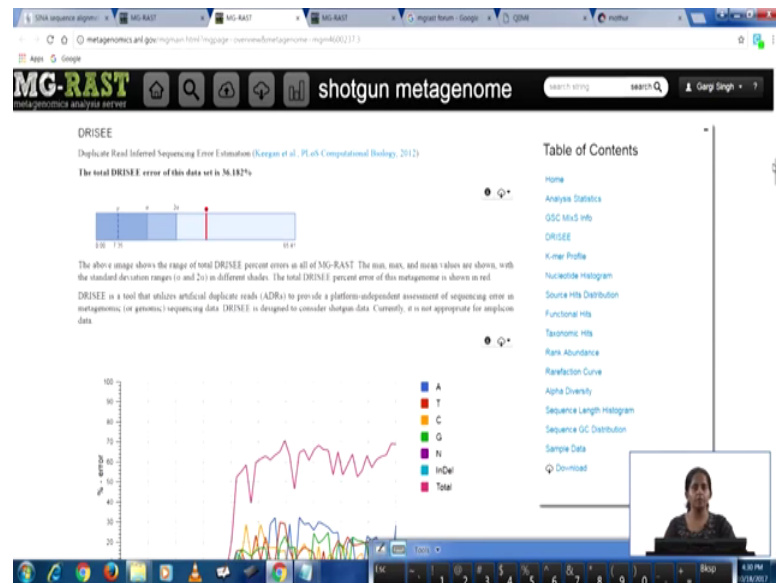
In that case it would be better for you to go for high throughput amplicon sequencing and focus just on this narrow strip of ribosomal RNA which is less than 1 percent.

(Refer Slide Time: 18:10)



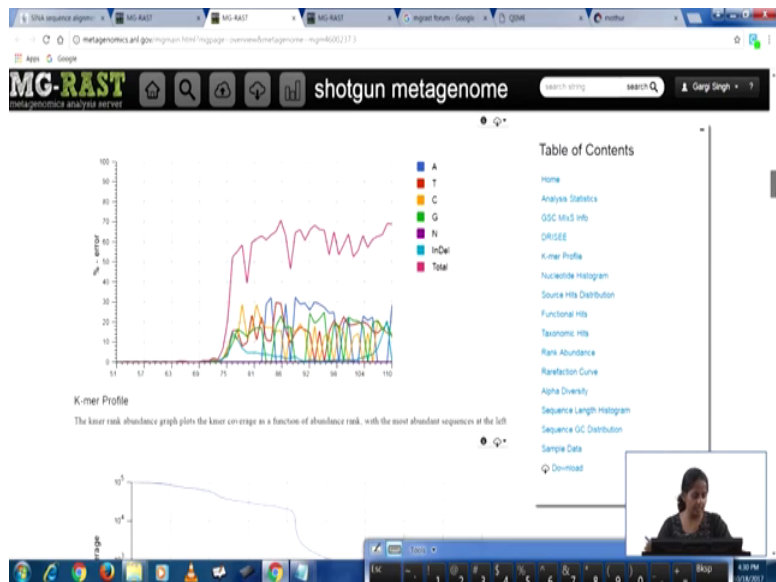
Then, here it will give you a basic analysis statistics like how many base pairs were uploaded how many sequences were uploaded and so on and so forth and then it does drisee analysis.

(Refer Slide Time: 18:19)



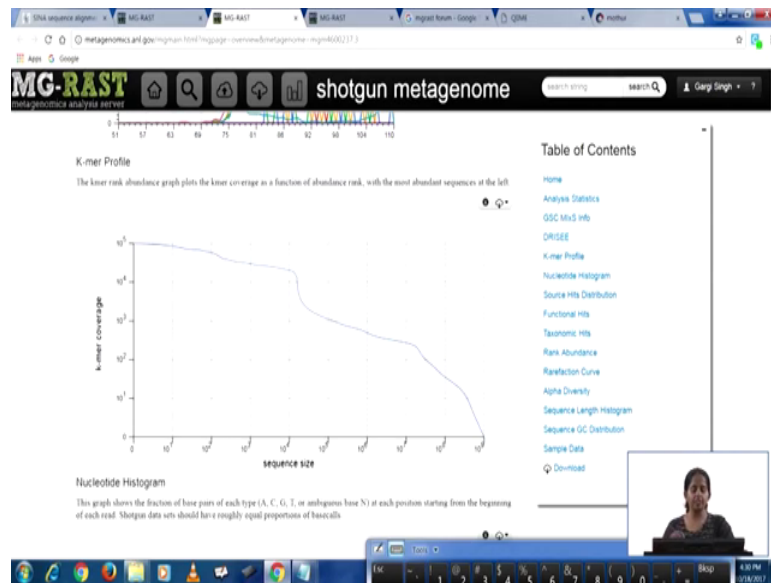
So, this drisee analysis will tell me about duplicate reads and where do I align the in the distribution of replicate reach.

(Refer Slide Time: 18:26)



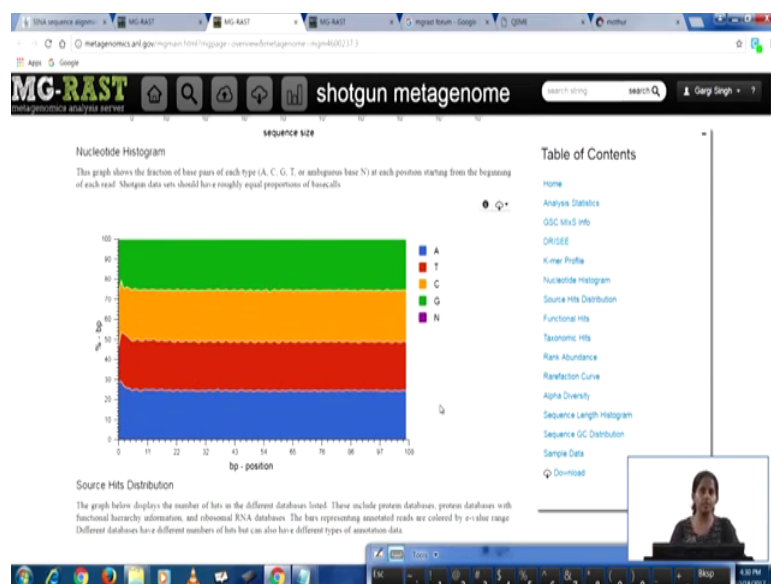
And then this is k mer profile k mer profile basically will tell how many times the abundant sequences appeared.

(Refer Slide Time: 18:28)



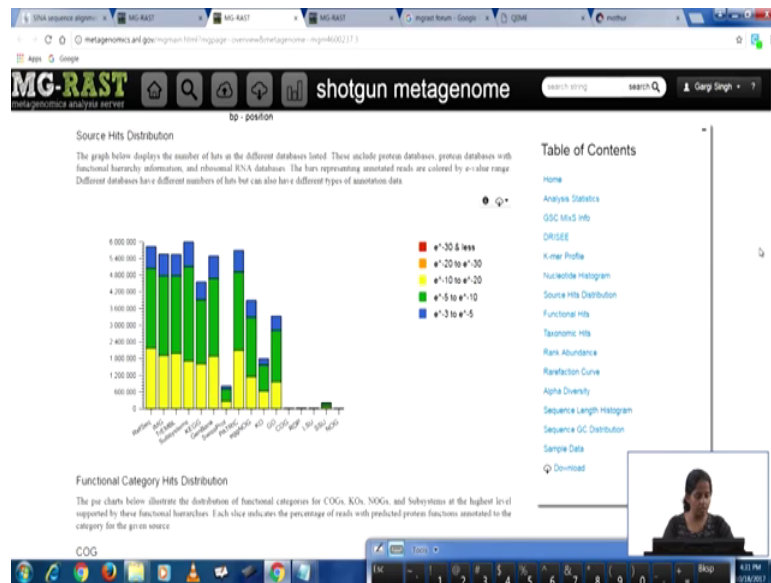
And then for most abundance we fall to less abundant.

(Refer Slide Time: 18:38)



And then this is nucleotide histogram nucleotide histogram gives a distinguishing distinguishes between the different kinds of nucleotides that are present in our sample for example, ATCG and N N empty nothing predicted and we can get an idea this is very important for us to know, because N is we do not know what it is it is ambiguous ok. And this is very important because G G C ratio gives us an idea of what kind of microbe it is microbes are there.

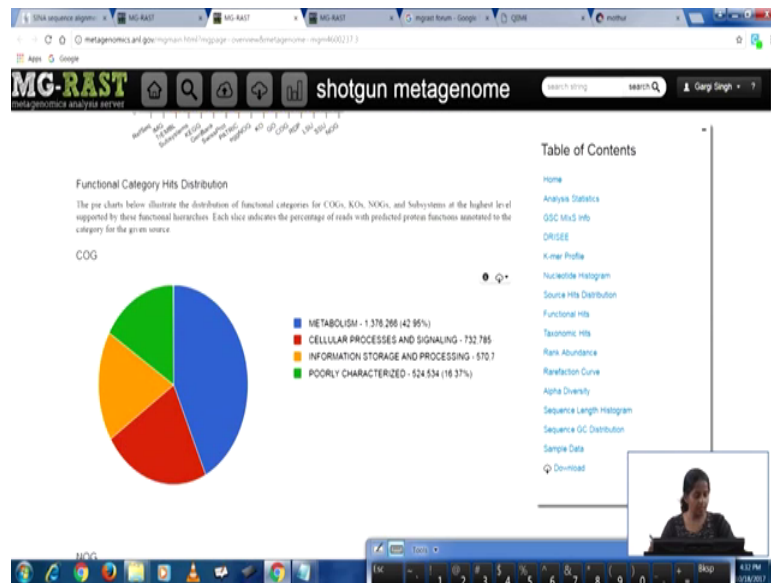
(Refer Slide Time: 19:06)



And, then source which distribution is after the quality checks were done on the same on the sequences that I uploaded, it matches them with different different databases and here are the databases and when it does that if we stand to give me an idea of how many reads were annotated.

So, if the e value is very low it is in red if event is very high it is in blue. And we notice that most of the reads happened were in subsystems ref seq and in patric and led not so, many in Swiss prot quite some in genbank not so, many in keg. So, these are all different tools different databases that are helpful for us in different perspectives genbank definitely for genes cake for metabolic pathways and so on and so forth RDP is a ribosomal database project were we talked about very important anyway.

(Refer Slide Time: 19:59)

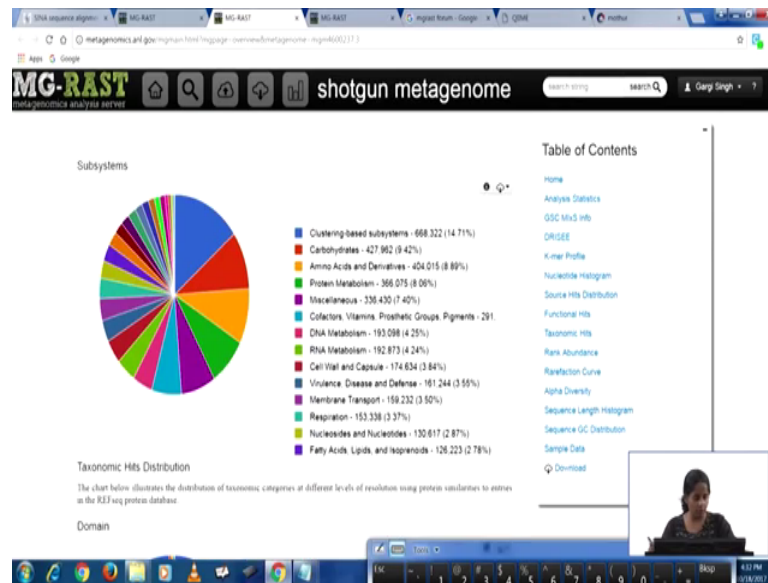


So, next it also tells me about functional category hits distribution. So, it will tell me already.

So, now this it will tell me for COGs, KOs and NOGs and Subsystems at highest-ever the function hierarchies. So, we know that 42 percent 42 nearly 43 percent were involved with metabolism and some 16 percent were not characterized properly our interval for information storage and processing the red were cellular processes and signalling.

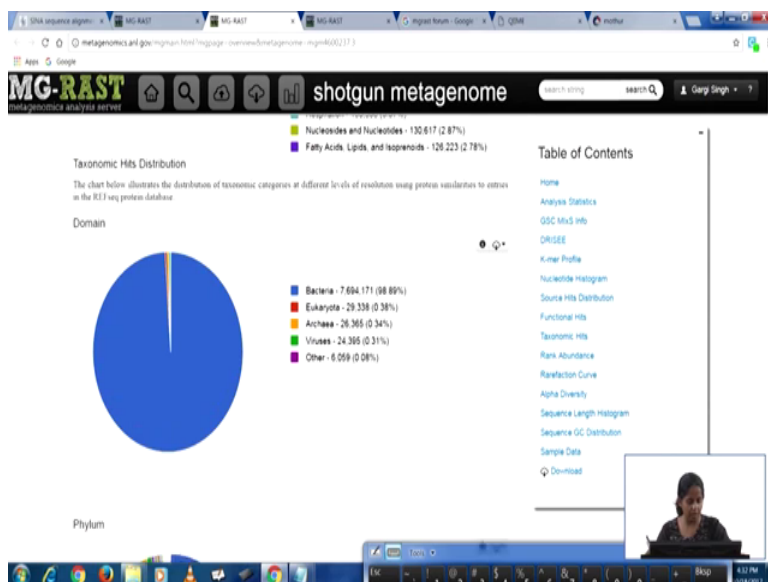
So, metabolism unless I am interested in certain functional traits were related to metabolism, this is housekeeping, this is housekeeping, this is also mostly housekeeping in this we do not know, all right when you look at NOG similar information this is more subsystem we have more detailed information.

(Refer Slide Time: 20:41)



So, most are clustering based subsystems and you have carbohydrate this is what my research was interested in by the way, then we have amino acids protein and so on and so forth.

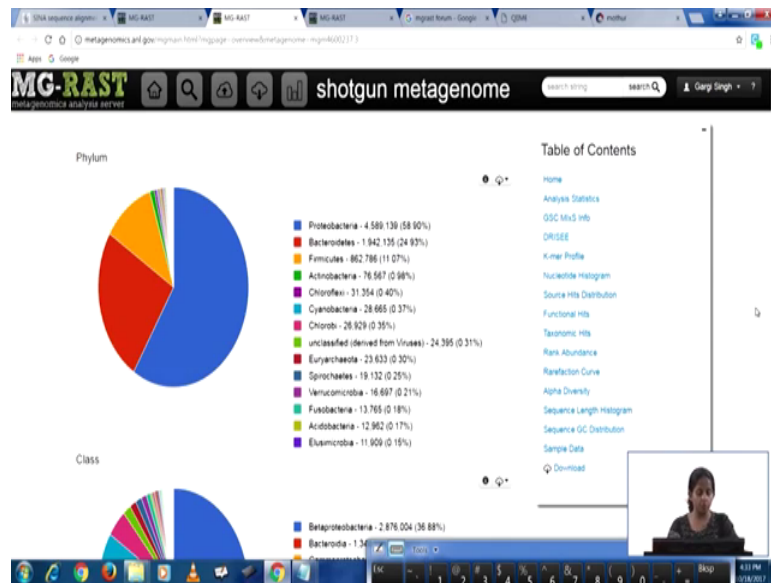
(Refer Slide Time: 20:56)



Then it can it will also inform me about taxonomic hits most of the sample most of sequences were bacterial sample 9 nearly 99 percent.

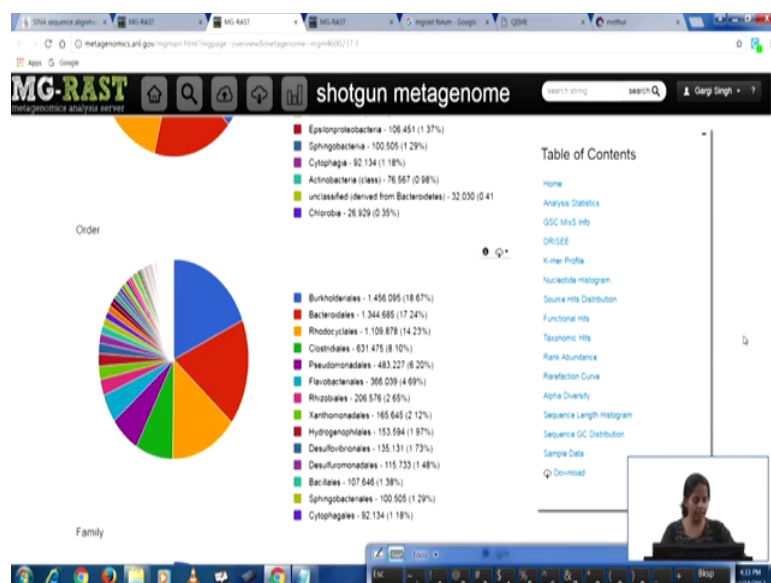
Very small portion of them were archaea, eukarya, archaea viruses and others.

(Refer Slide Time: 21:13)



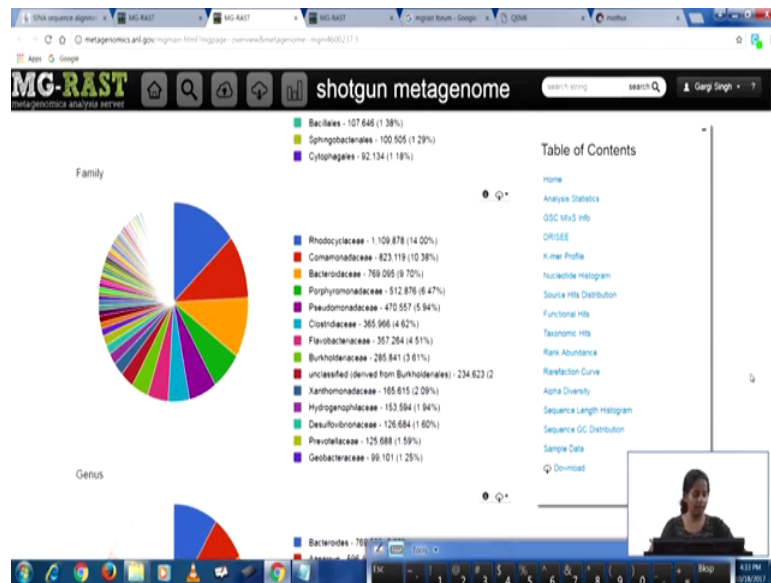
So, if I go for on phylum level I see that most of them are proteobacteria, some are Bacteroidetes, some are firmicutes I was very interested in bacteroidetes and firmicutes and in actinobacteria chloroflexi so on and so forth. And then on class level most of them were beta proteobacteria followed by bacteroidia, followed by gamma proteobacteria, followed by clostridia, I was very interested in clostridia and in bacteroidia.

(Refer Slide Time: 21:39)



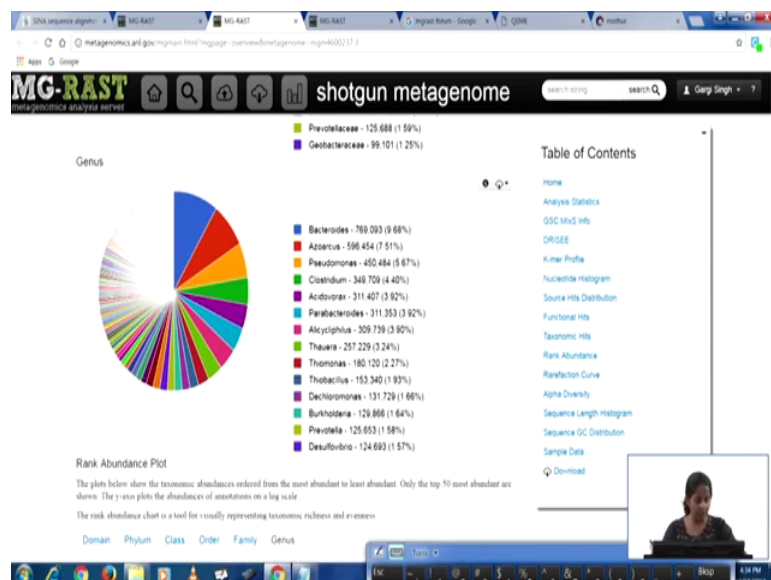
Then in order level again it divides into different different parts again clostridia was very important because this was a cellulose degrading study.

(Refer Slide Time: 21:47)



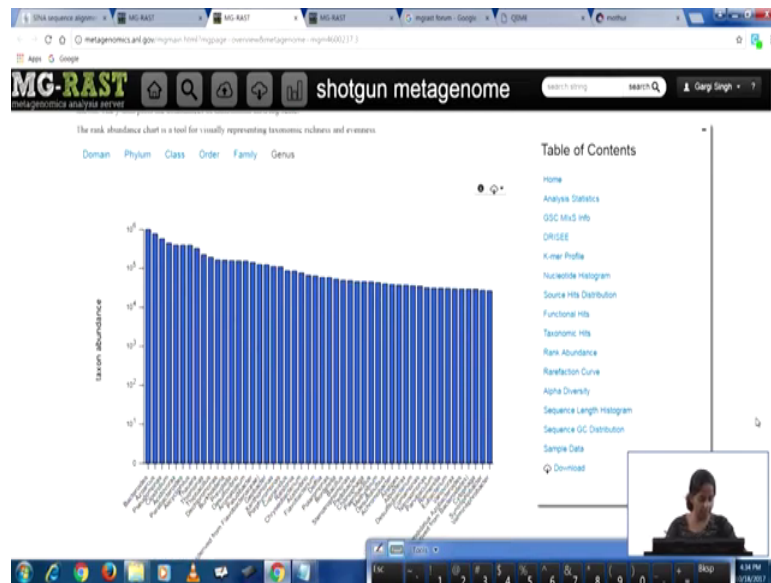
And, then it will also tell me on family level. Now look at these plots it is very hard to tell what the data is just by looking at these pie charts, but you can actually download the data here and once you download the data you can actually draw your own diagrams and do your own analyses.

(Refer Slide Time: 22:02)



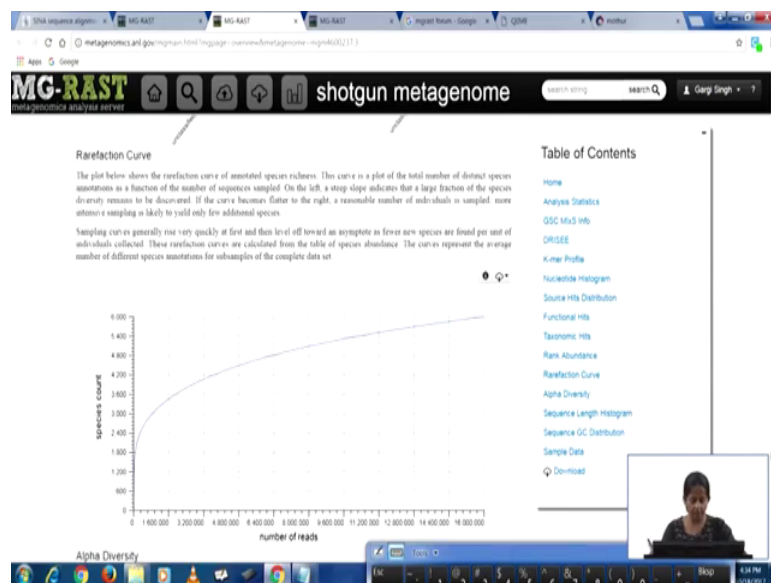
So, this is a beauty of MG RAST it analyses things for you and this is on genus level all rightly.

(Refer Slide Time: 22:06)



And, then this is rank abundance plot. So, this is an abundance of the taxon and this is which toxin we're talking about a lot of bacteria least amount of vermin affair of bacteria.

(Refer Slide Time: 22:16)

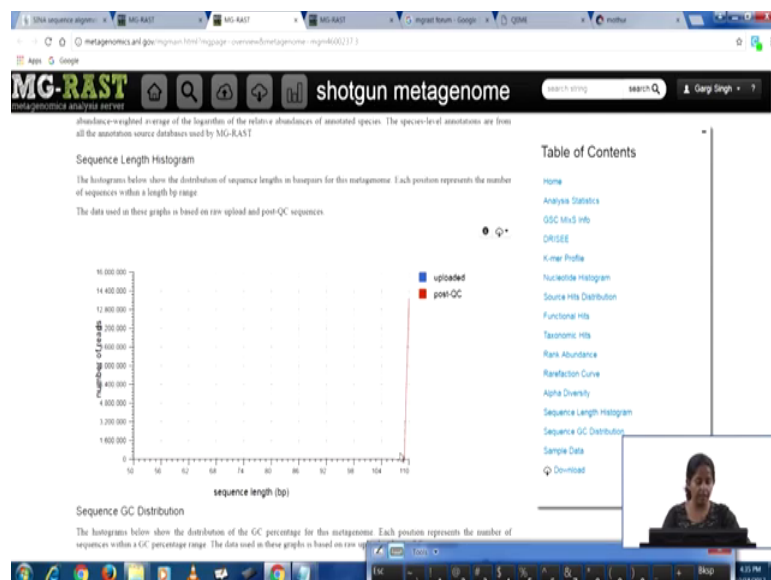


And, then this is rarefaction curve. So, if you look at a rarefaction curve rarefaction curve is also called collectors curve, what it tells me is that was my sampling depth in if sufficient or not. So, if you look at my career rarefaction curve it has not plateaued, which means that a number of new sequence, new species, that we could have detected was if we had more sequences, if he had more sequencing depth would be higher.

So, we had not we have not this sample this data is not completely representative of all microbial members in the community. This is very important curve and this is very important information. So, if I had like this then I have to exceed the limitation in whatever on conclusions that I make that this data is perhaps not very good representation of what is really happening, because the collectors curve or the rarefaction curve has not plateaued.

So, once it starts tattooing once it plateaus we do not detect any more species no matter how much you increase the sequencing gap?

(Refer Slide Time: 23:17)



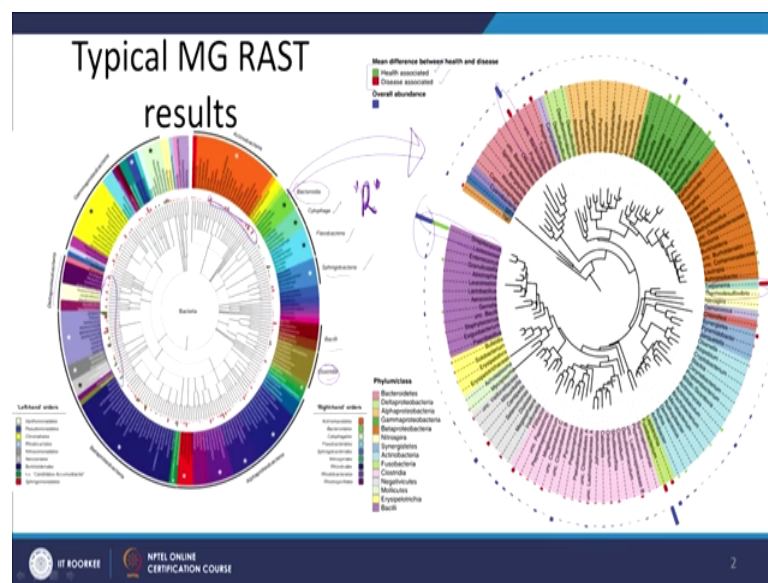
So, basically you have reached the optimal sequencing depth, then we have alpha diversity and then we also have sequence length dendrogram most of them were around 100 and 8. So, this is good this is the target length of sequences this also tells me GC distribution it gives me more information about sample people who had submitted the sample and so forth.

So, for each sample you will get this information in for basic analysis this is enough all righty it is also possible that we can compare two samples and in this is where the beauty appears. So, in MG RAST once you have uploaded your data and it might take some time just the minute act of uploading the data, you got to either decide to make it publicly available or make the data private.

Now, for example, the data that I was uploading in this lecture earlier is private data. So, now, the issue with private data is that if you data is private; obviously, the benefit is that no 1 can view it and use your data and publish results, but the issue with that is that that algorithm the processing MG RAST takes a long time, because it is given the least priority.

Whereas, if you make your data public you are likely to receive data analyses in man maybe a couple of days or maybe maximum couple of weeks. So, here I want to give you a glimpse of how MG RAST results may look like in previous examples, I showed you per and per sample analysis of the meta-genome that you are uploaded on MG RAST here let us take a look about, how we can use the data for MG RAST to make comparative analysis between different samples so, here on the left panel here.

(Refer Slide Time: 24:53)



We have this particular dendrogram from one sample and even though this is from MG rest by the way.

And you notice that MG RAST can give you information on class level phylum level or family level species level you decide the level of resolution that you want your data to be represented. And these dendrogram are very very helpful for us to understand what is the diversity of your sample and how many different kinds of microbes are present and this helps us also understand, what kind of functions might be happening because for example, if clostridia is present in large quantities, we can suspect that maybe this is a

loss degrading environment or if delta proteobacteria are present in large numbers we can suspect sulphate reducing environment.

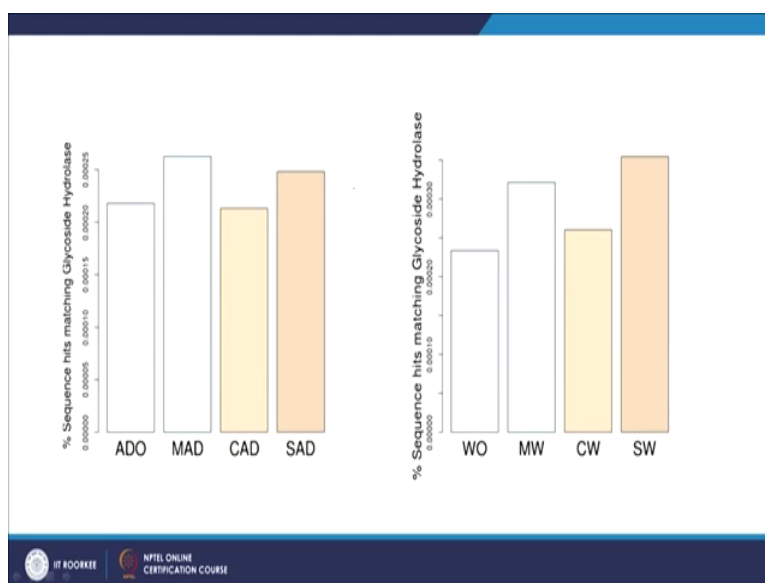
Another thing about MG RAST is that not only does it give you a dendrogram that informs you about which of the microbes are present, but it also gives you an idea of in how much abundance. So, if you can find out these little bars here in your data they will give you idea of relative abundance of the sequences that match the particular species class or family or phylum.

Now, this information can be used to generate another kinds of plot which is on the right panel here, where you see not only can you see the different kinds of microbes present which species which classes are present here, but now you can make comparative analysis. For example, there are 2 different kinds of samples being analysed here one is from a healthy some healthy mammal in the others from diseased mammal. And we can compare which microbes are present more in healthy animal versus in diseased animal this is a very good visual image that has been made using data that can be made for using data from MG RAST.

So, for example, we can clearly see that streptococcus was present an high number in healthy mammal whereas, Treponema and Prevotella were more present and unhealthy mammal. And not only that it will you can also get the information on what is the overall abundance, which is given by the blue circle it is actually not a circle, but it is the layout of blue bars and this is one example of how you can use MG RAST data to make really good visual representations of your data.

Now, there many packages on software are which I have recommended earlier in this class and I highly recommend you to explore these packages, that you can use to draw wonderful diagrams and make your data more clear and more accessible for scientific and general public.

(Refer Slide Time: 27:23)



Now, here is the other data that I had uploaded in this lecture earlier here is some analyses from that data. So, not only can you actually download the figures made by MG RAST itself for example, the one on the left, but you what you can do is you can download data at every step of processing analyses and then you can make your own graphs to make your own analyses.

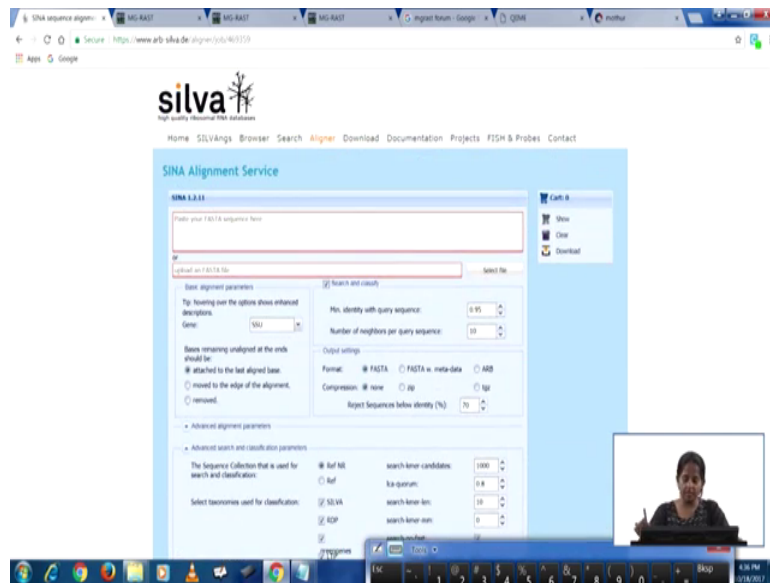
So, these are 8 different samples that were uploaded on MG RAST and analysed for my nanocellulose project. And I was a very interested in glycoside hydrolase family of enzymes and now here I can I add a downloaded the data and found the enzymes that are that belong to glycoside hydrolase family and have similar functions or functions that I am interested in namely degradation of beta glycosides bond.

So, now when I collected a number of sequences in each of these in these families, I could draw these bar plots that help me understand that helped me make comparative analysis from different samples. For example, this sample has been incubated for more than 200 days it was taken from a wetland it is your original sample as an original for the batch I say and then it was exposed to microbial cellulose cationic nanocellulose. And sulfur sulfuric acid reduced and ionic nanocellulose and I can see how after a period of time which is 4 weeks how much the glycoside hydrolase have changed from the initial inocula over time.

Now, similarly this graph here is also very informative. So, the beauty of MG RAST is that if you find it challenging to analyse data or make visual representations of data there are sufficient tools available on the online platform itself for you to make very pretty diagrams meaningful diagrams.

And you also have the flexibility to download your data and interpret it and represent it in whatever way suits that suit you and your paper.

(Refer Slide Time: 29:20)



Let us switch back to sina so, remember we used silvas sina to see to align our 2 16 S RRNA sanger sequences, it has finished them and it tells me that the first one found they found 95 percentage similarity, in the second they found 91 parentage similarity so, let us display their classification.

(Refer Slide Time: 29:43)

The screenshot shows the SILVA Incremental Aligner (SINA) web interface. At the top, there are search filters for taxonomic levels: SILVA, RDP, TIGR, and EMBL. Below these, a table titled 'Aligner Task Manager' shows a job named 'query1' with a status of 'Finished'. The 'Alignment Result Table' displays two results, both identified as 'Bacteria Proteobacteria Gammaproteobacteria Gammaproteobacteria uncultured'. The interface includes a 'Download Results' button and a 'Export to CSV' option. A small video inset in the bottom right corner shows a person speaking.

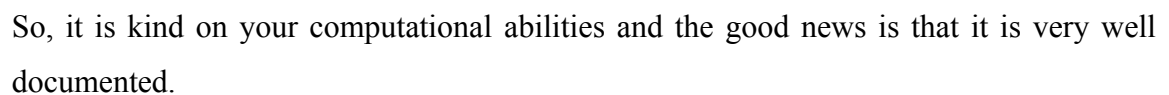
The first one it said is gamma proteobacteria, but it did not stop there it goes even ahead and says what kind of gamma proteobacteria we talking about here. It says it is (Refer Time: 29:47) and the second one is (Refer Time: 29:50) cool and I also guarantee and I can actually export it to CSV file and then I can get more further data that I need.

So, student is this is sina for you, now when it comes to high throughput amplicon sequencing we do not need to go to MG RAST. We can use another other we use other platforms and the 2 leading platforms as I mentioned earlier are qiime and mothur.

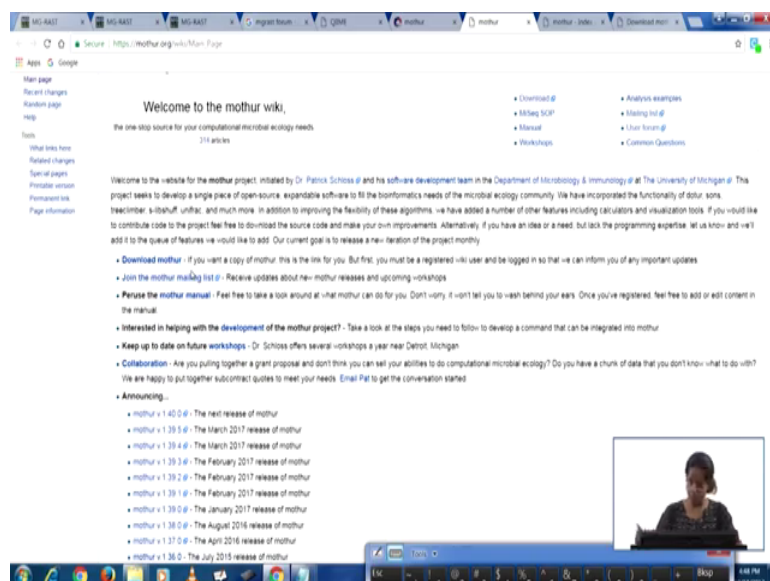
(Refer Slide Time: 30:18)

The screenshot shows the mothur project website. The header includes the 'mothur' logo and navigation links for 'Download', 'Wiki', 'Forum', and 'facebook'. The main content area welcomes visitors to the mothur project, initiated by Dr. Frank Schloss and his software development team at the Department of Microbiology & Immunology at The University of Michigan. It describes mothur as an open-source, extensible software for the bioinformatics needs of the microbial ecology community. A 'Subscribe to mothur mailing list' section is visible, along with a 'Download' button. The footer mentions the Department of Microbiology & Immunology at The University of Michigan Medical School and The University of Michigan, with a copyright notice for 2008-2017. A small video inset in the bottom right corner shows a person speaking.

(Refer Slide Time: 30:49)

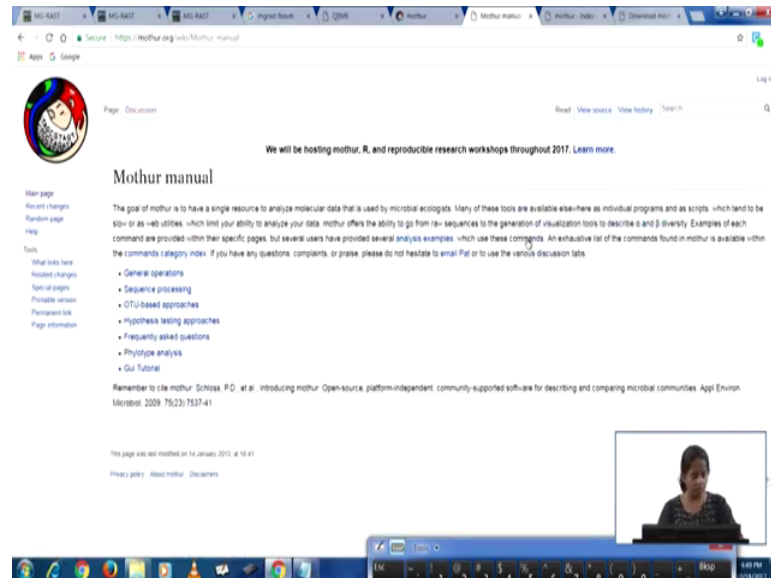


(Refer Slide Time: 30:55)



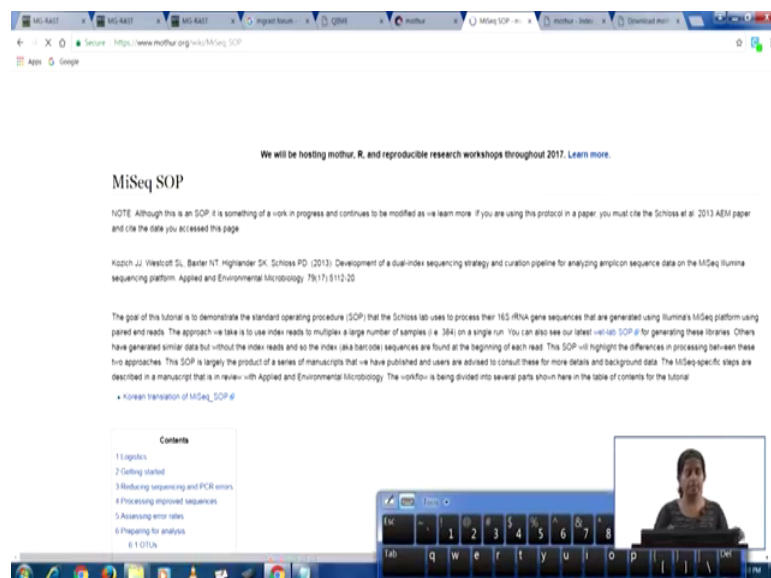
So, this is the wiki page of mothur and if you are interested you can look at how to download mother, how to join it is mailing this and so forth and so on.

(Refer Slide Time: 31:06)



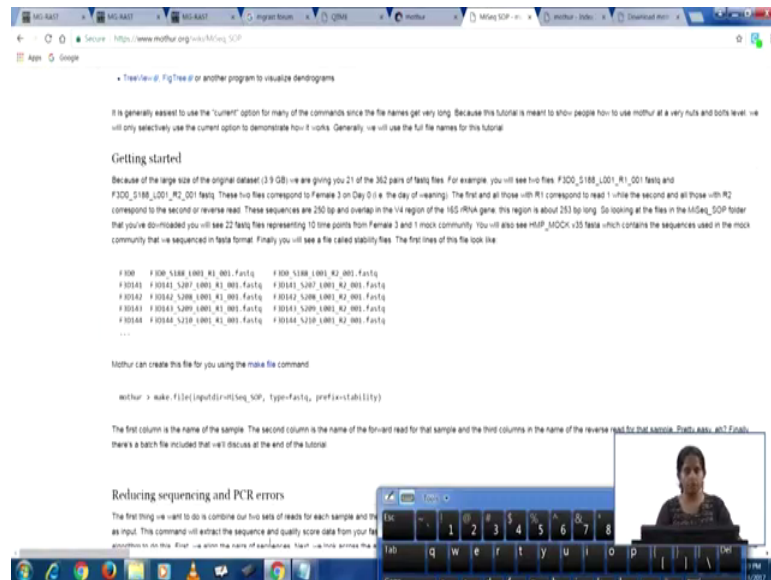
And you also can have manuals. So, let us to get manuals. So, in manuals what you can do is you can go through different different kinds of operations that you are interested in general operation sequence processing operation and so on. And find out what the commands are. So, let us look at one particular operation that is very commonly used mothur 16 S analysis we came.

(Refer Slide Time: 31:43)



So, this is the MiSeq SOP for mothur. So, in this what they have; obviously, they had the citation here and now if you know go through this they will give you step-by-step information on how to use mother?

(Refer Slide Time: 31:54)



This is the data the mock data that they are using.

And these are the files in the mock data. So, these are comparisons. For example, this perhaps the one paired one end of the pair and there is the other end the paired and. So, these are the paired for readings. So, we need attach them we get one long sequence. So, you make 1 stability file that will call your first few files, and then the next step would be you want to remove that PCR errors.

(Refer Slide Time: 32:31)

and you would get the same output for each. For the purposes of this tutorial we will write out the names of the files. At this point our sequencing error rate has probably dropped more than an order of magnitude and we have 128872 sequences. Let's press on.

Processing improved sequences

We anticipate that many of our sequences are duplicates of each other. Because it's computationally wasteful to align the same thing a billion times, we'll unique our sequences using the `unique seqs` command.

```
author > unique_seqs(fasta=stability.trim.config.good.fasta)
```

If two sequences have the same identical sequence, then they're considered duplicates and will get merged. In the screen output there are two columns - the first is the number of sequences characterized and the second is the number of unique sequences remaining. So after running unique seqs we have gone from 128872 to 16426 sequences. This will make our life much easier. Another thing to do to make our lives easier is to simplify the names and group files. If you look at the most recent versions of those files you'll see together they are 13 kb. This may not seem like much, but with a full 16S run those long sequence names can add up and make life tedious. So we'll run `count_seqs` to generate a table where the rows are the names of the unique sequences and the columns are the names of the groups. The table is then filled with the number of times each unique sequence shows up in each group.

```
author > count_seqs(names=stability.trim.config.good.names, group=stability.config.good.groups)
```

This will generate a file called `stability.trim.config.good.count_table`. In subsequent commands we'll use it by using the `count` option.

```
author > summary_seqs(count=stability.trim.config.good.count_table)
```

Using `stability.trim.config.good.unique.fasta` as input file for the `fasta` parameter.

using 8 processors.

	Start	End	Mbases	Abigs	Polymer	Isotigs
Minimum:	1	250	250	0	3	1
J-SS-111a:	1	252	252	0	1	1222
ZS-111a:	1	252	252	0	4	32219
Median:	1	252	252	0	4	68837
J-SS-111a:	1	253	253	0	5	98655
J-SS-111a:	1	253	253	0	6	125651
Maximum:	1	270	270	0	12	128872
Mean:	1	252.462	252.462	0	4.36693	
# of unique seqs:						16426
total # of seqs:						128872

So, they are making removing the PCR errors here and then they process improved sequences. So, you want to find out what are the unique sequences you want to remove other sequences, if you do count sequences count on 6.

(Refer Slide Time: 32:39)

```
author > count_seqs(names=stability.trim.config.good.names, group=stability.config.good.groups)
```

This will generate a file called `stability.trim.config.good.count_table`. In subsequent commands we'll use it by using the `count` option.

```
author > summary_seqs(count=stability.trim.config.good.count_table)
```

Using `stability.trim.config.good.unique.fasta` as input file for the `fasta` parameter.

using 8 processors.

	Start	End	Mbases	Abigs	Polymer	Isotigs
Minimum:	1	250	250	0	3	1
J-SS-111a:	1	252	252	0	1	1222
ZS-111a:	1	252	252	0	4	32219
Median:	1	252	252	0	4	68837
J-SS-111a:	1	253	253	0	5	98655
J-SS-111a:	1	253	253	0	6	125651
Maximum:	1	270	270	0	12	128872
Mean:	1	252.462	252.462	0	4.36693	
# of unique seqs:						16426
total # of seqs:						128872

Cool, right? Now we need to align our sequences to the reference alignment. Again we can make our lives a bit easier by making a database customized to our region of interest using the `pcr_seqs` command. To run this command you need to have the reference database (`silva.bacteria.fasta`) and know where in that alignment your sequences start and end. To remove the leading and trailing dots we will keep only to follow. You could also run this command using your primers of interest.

```
author > pcr_seqs(fasta=silva.bacteria.fasta, start=11894, end=2319, keepdots=f, processors=8)
```

Let's rename it to something more useful using the `rename` file command.

```
author > rename.file(input=silva.bacteria.pcr.fasta, new=silva.v4)
```

Let's take a look at what we've made

And then summary dot sequences summary dot anything you can always get a summary.

(Refer Slide Time: 32:43)

[illegible]

And, then you can rename the file and so on and so forth in the alignment option comes here.

(Refer Slide Time: 32:52)

The screenshot shows a terminal window with several commands and their outputs:

```
# of seqs: 10056
```

Now we have a customized reference alignment to align our sequences! The nice thing about this reference is that instead of being 50,000 columns wide, it is now 13,425 columns wide which will save our hard drive some space and should improve the overall alignment quality. We'll do the alignment with `align_seqs`:

```
mother > align_seqs(fasta=stability.trix.config.god.unique.fasta, reference=cvlva.st.fasta)
```

This should be done in a manner of seconds and we can run `summary_seqs` again:

```
mother > summary_seqs(fasta=stability.trix.config.god.unique.align, count=stability.trix.config.god.count_table)
using 8 processors.
```

	Start	End	Hitses	Ambigs	Polymer	Mutations
Minimum:	1250	11550	250	0	1	1
2.5% tile:	1968	11550	252	0	1	3222
75% tile:	1968	11550	252	0	4	32229
Median:	1968	11550	252	0	4	64417
75% tile:	1968	11550	251	0	5	95605
97.5% tile:	1968	11550	251	0	8	125613
Maximum:	1967	11549	250	0	12	128872
Mean:	1967.99	11550	252	462	4	36693
# of unique seqs:			16426			
Total # of seqs:			128872			

So what does this mean? You'll see that the bulk of the sequences start at position 1968 and end at position 11550. Some sequences start at position 1250 or 196 from the mode positions are likely due to an insertion or deletion at the terminus ends of the alignments. Sometimes you'll see sequences that start and end at the which is generally due to non-specific amplification. To make sure that everything overlaps the same region we'll re-run screen-seqs to get sequences that start at position 11550. (We'll also set the maximum homopolymer length to 8 since there's no execution of screen-seqs above). Note that we need the count table so that we can do at the start and stop positions.

And, when you align you have the option to choose the database you want to choose here they are using silva, but if you want to use another database you can use that and then here is a summary.

(Refer Slide Time: 33:03)

```
MS-AASIT > MS-AASIT > QXONE > MSAFSCP -xw -m ...> Downloaded successfully
```

← 🔍 🔄 ⌂ Secure https://www.mothur.org/wiki/MoQing_SCP

```
> Apps Google
```

```
length of the original alignment : 13425  
Number of sequences used to construct filter: 16286
```

This means that our initial alignment was 13425 columns wide and that we were able to remove 13049 terminal gap characters using truncate and vertical gap characters using vertcat=1. The final alignment length is 376 columns. Because we've perhaps created some redundancy across our sequences by trimming the ends, we can run unique_seqs:

```
mothur > unique.seqs(fasta=stability.trin.config.good.unique.good.filter.fasta, count=stability.trin.config.good.count_table)
```

This identified 3 duplicate sequences that we've now merged with previous unique sequences. The next thing we want to do to further de-noise our sequences is to pre-cluster the sequences using the pr.clust command allowing us up to 2 differences between sequences by group and then sort them by abundance and go from most abundant to least and identify sequence clusters as they are either identical or if they are then they get merged. We generally favor assigning 1 difference for every 100 bp of sequence.

```
mothur > pr.cluster(fasta=stability.trin.config.good.unique.good.filter.unique.fasta, count=stability.trin.config.good.count_table, diffs=2)
```

And, then you can filter your sequences and you again find you need then you can click you can cluster your sequences. So, for clustering you do pre clustering then you look for chimera and then you remove the chimera sequences that you got from chimera dot research. And then you can classify your sequences you can remove certain lineages that you know should not be present in the sample anyway.

(Refer Slide Time: 33:34)

sequences, we'd have 34 OTUs from the Chocoma community. This number of course includes some stealthy chimeras that escaped our detection method. If we used 3000 sequences, we would have about 31 OTUs in a perfect world with no chimeras and no sequencing errors. We'd have 20 OTUs. This is not a perfectly good. This is a pretty damn good.

Preparing for analysis

We're almost to the point where you can have some fun with your data (I'm already having fun, aren't you?). We'd like to do two things: assign sequences to OTUs and phyloTypes. First, we want to remove the block sample from our dataset using the [remove groups](#) command:

```
mkdir -p remove_group(count-stability.trim.config.good.unique.god.filter.unique.precluster.denovo.vsearch.pick.pick_count_table, fast-stability.trim.config.good.unique.god.filter.unique.precluster.denovo.vsearch.pick.pick_count_table)
```

OTUs

Now we have a couple of options for clustering sequences into OTUs. For a small dataset like this, we can do the traditional approach using `dist.sep` and `cluster`:

```
mkdir -p dist.sep(fast-stability.trim.config.good.unique.god.filter.unique.precluster.pick.pick_count_table, cutoff=0)
mkdir -p cluster(column-stability.trim.config.good.unique.god.filter.unique.precluster.pick.pick_count_table, count-stability.trim.config.good.unique.god.filter.unique.precluster.pick.pick_count_table)
```

Clustering stability

iter	time	label	num. otus	cutoff	fp	fn	fp	fn	sensitivity	specificity	ppo	npv	fdr	accuracy	mcc	f-score
0	0	0.01	2236	0.91	0	2453489	0	21336	0	0.99385	0	0.99385	0	0.99385	0	0.99385
1	0	0.01	509	0.91	13444	2453489	1380	34752	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385
2	0	0.01	437	0.91	14273	2453489	1714	2884	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385
3	0	0.01	437	0.91	14273	2453489	1707	2884	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385
4	0	0.01	437	0.91	14273	2453489	1701	2884	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385
5	0	0.01	437	0.91	14273	2453489	1701	2884	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385	0.99385

The alternative to using `cluster` (command) is in this approach, we use the taxonomic information to split the sequences into bins and then cluster within each bin. The results of the clustering process are shown in the table above. The clustering process is successful, with high accuracy and low error rates. A small inset video shows a person speaking, likely the instructor.

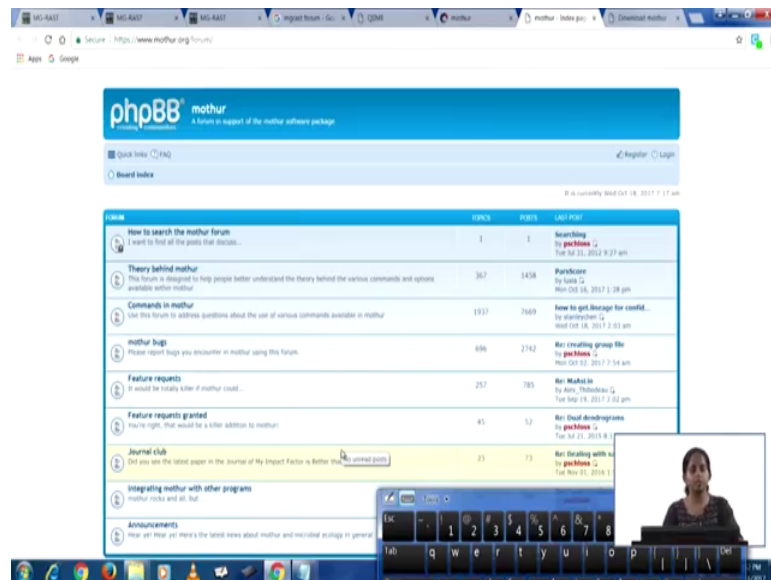
And then finally, you will get your final data which in which you can find out what is your error rate in this particular case and then you can analyse you can find their

So, you will never be lost and there is almost all kind of data analysis you need to do there inform you all right.

[illegible]

So, we had been clustering we can do classification and then you can do for you can do phylotypes analyses, you can classify you can also do phylogenetic analyses and see this entirely wiki page will help you a lot in trying to do even if I calculating alpha diversity beta diversity and so on and so forth. So, this is my seek.

(Refer Slide Time: 34:29)



And, if there are questions that you face when you are not able to find it when you're trying to follow a particular SOP on from the (Refer Time: 34:36) of mother, then there is this wonderful forum here which is which supports some mothur software package.

So, you have a lot of questions and you can question if it there is a feature you want to see you can feature that in here. And this is a very supportive community and very quickly someone will respond and help you with your analyses all righty students this is mothur for you.

(Refer Slide Time: 35:04)



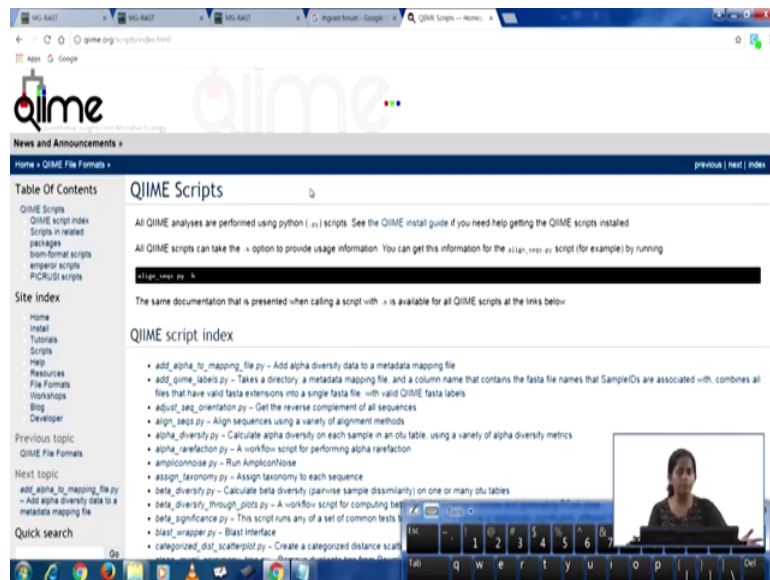
The next we have is qiime. So, in qiime, it is quantitative insight into microbial ecology, in qiime now they are moving to qiime too which is good, but in qiime it requires a platform if the computational needs are much higher than that of mothur and in qiime if you are using a windows or mac os or linux you might want to install the virtual box, and that is the easiest way or you can actually you can use virtual box to install qiime or you have even installation of qiime is a challenge actually or you will have to go through make sure dependencies are present and then you can install it and then run it.

(Refer Slide Time: 35:38)



Similarly, like mothur they have very good script documentation that will help you, do your analysis and again this will do analysis for all platforms.

(Refer Slide Time: 35:46)



So, that is part of sequencing alumina based sequencing or ion torrent based sequencing or applied bio system based sequencing whatever kind of sequencing they be lot these are all commands that you can use. One difference that I have noted between qiime and mothur is mothur is computationally kinder than qiime.

Qiime tends to make prettier plots because data visualization is so, important. The other thing is that certain funding agencies now required people to submit their data to qiime in to use time in US. So, in that case qiime does have an upper hand, but mothur offers lot of flexibility. So, if you know coding if you're hardcore bioinformatician I see that there are advantages and disadvantages for both mothur and qiime both are equally good and both are equally useful.

Dear students, I hope that this lecture and the previous lecture gave you valuable real information on how you can use tools that are easily available to you, to analyse your sequences and make sense out of them and that this we conclude our lecture series on our course applied environmental microbiology; All the best for all your assignments and home exams.

Thank you.