So what I want to do today is recap a little bit what we talked about last time, reiterate some of the important points, and then show you how we can learn something about microorganisms in the environment by talking about in-situ identification of microorganisms as well as genomics. We'll first talk about genomics and general and then talk about some applications of genomics to environmental microbiology because I think there is some of the most exciting new developments are in the area, actually. So last time we talked about molecular evolution and ecology. And just to recap, some of the main points were that we can actually use genes or gene sequences for a couple of very important questions that we want to explore. The first one was gene sequences act as evolutionary chronometers. Now, what do I mean by that? Basically what we said last time was that each gene, each sequence in the genome accumulates mutations with a certain probability. So what we mean is that all genes accumulate mutations over time. Now, these of course are the mutations that do not kill the organisms, so not the deleterious mutations, but these are mutations that are either slightly deleterious, or don't matter, or are beneficial mutations. OK, and what this entails is that each gene accumulates mutation with a certain probability over time. It basically means that two organisms that come from species that are relatively closely related to each other have gene sequences that will be much more similar to each other than genes from an organism that comes from a species that's much more distantly related. So, in practical terms what this means is your genes are much, much more similar to those of a monkey than they are to a crocodile, for example. And we can take advantage of that by applying some algorithms, some mathematical modeling essentially, to constrain these relationships in those phylogenetic trees that we talked about last time. And I also mentioned that the ribosomal RNA genes are particularly important for that process. In principle, you could do it with any protein coding machine or any kind of gene in the genome, but we use the ribosomal RNA genes in particular because all organisms have them. They're part of a handful of genes that are what we called universally distributed last time. And what this allows us to do is then construct phylogenetic relationships for all living organisms. And I just want to remind you of the tree of life that I showed you last time where we can really explore the relationships amongst all living organisms. And some of the important points there that we made were, for example, that the tree of life supports the endosymbiont theory, that when you actually look on the tree where the mitochondria and the chloroplasts tree, they fall into the bacteria. Now, there is a question where somebody asked in the online survey, can the mitochondria and chloroplasts still live outside of the eukaryotic cell? And the answer is no, they can't anymore because over evolutionary time the two organisms have become so integrated that the mitochondria and chloroplasts both lost their ability to live outside of the eukaryotic host cell. Another important point that we made last time that I want to reiterate here is that gene sequences, when we go into the environment, and obtain them directly from the environment act as a proxy for microbial diversity in the environment. So, the number of genes recovered directly from the environment is a measure of diversity. And we said that this actually plays a very, very important role in the analysis of microbial communities, and I showed you the example here where we went and took some ocean water and basically apply this technique that outlined last time where we can actually amplify ribosomal RNA genes from environmental samples, clone them, determine the sequence, and then constructs phylogenetic trees. And what you see here is a tree where we summarize the major groups that we found in the sample have been only for two of those groups where we show the entire set of sequences that we actually obtained because there were so many of them out there. And what we basically found was that over 1500 bacterial 16S ribosomal RNA gene sequences coexist in this environment. And what we said also last time was that the analyses like these have really taught us that microorganisms are the most diverse organisms on the planet. So, most diversity is amongst the microorganisms, and one of the big questions now is what are all those microorganisms doing in the environment? And so, today what I want to do with you is basically explore this question of how we can actually figure out what those microorganisms are all doing in environmental samples? So we can say we are exploring the function of microbes in the environment. At first, I want to cover how we can actually identify them in the environment. And I want to show you one specific example, and then I want to talk about genomics in general, and then basically end with an application of genomics to environmental questions. So, let's first talk about the in-situ identification of microorganisms. And the basic problem that I alluded to already before is that most microbes are only known [SOUND OFF/THEN ON] -- from 16S ribosomal RNA clone libraries. And we basically want to search and identify them in the environment. OK, and I'll show you a specific example of that later on. Now last time, we said that the ribosomal RNA sequences consist really, like all gene sequences, in fact. We identified several stretches of nucleotides, types of stretches, that can be found. We said the A type stretches and B type stretches that are very important for construction of phylogenetic relationships, because we can align them and look for changes in the nucleotide sequences because they are the same length and only differ in mutation and single nucleotide base pair changes. But then there's also those C type stretches, if you remember, and those we said vary at much faster rates because they are not functionally constrained in those genes. OK, so they can actually also accumulate length changes. And, it's these C type stretches that we can use sort of as diagnostic sequence stretches for microorganisms. So, what we can say is we identify organisms by the C type stretches, C type sequence stretches. And we call those signature sequences. OK, and they allow the differentiation of closely related organisms, -- because they vary at very fast rates between organisms. And the way we do this is that we construct so-called phylogenetic probes. I should probably write this over here. Now what are those phylogenetic probes? They're basically short pieces of DNA that have a fluorescent molecule attached to them. -- DNA molecules that are roughly 20 nucleotides in length, and they carry a florescent molecule. Now what the short, single-stranded stretches of DNA basically are is they are complementary to those C type sequence stretches -- -- in the ribosomal RNA. And so basically what we can do is we can collect microbial cells from the environment -- -- make them permeable -- -- and then basically mix them with those phylogenetic probes. And these probes will then permeate into the cell and bind to their complementary sequences. Then we wash away the unbound probe -

then permeate into the cell and bind to their complementary sequences. Then we wash away the unbound probe - -- and we can view it in a microscope under UV light. Let me show you an example of this. What you see here is basically a light micrograph. So this is what you see basically when you collect microbial cells from the environment under the microscope. Most bacteria look the same, so you cannot actually differentiate them all by just looking at them. But then these cells were fixed and permeabilized and then basically mixed with two different phylogenetic probes that identified two different types of organisms. One was labeled with a red Fluor, the other one with a green Fluor. And what you see is that you can now differentiate those two organisms. Now, why is this especially interesting? Well here's just a specific example where people were looking for bacteria capable of nitrogen oxidation. These are bacteria that are very important in, for example, sewage treatment. And it was known that there were two different types out there, one that oxidizes ammonia to nitrite, -- and that a second one that oxidizes nitrite to nitrate. And by doing this type of analysis what people basically learned is that those two organisms live in very, very close proximity at all times. So the organisms that oxidized ammonia to nitrite are really attached, and oftentimes even surround by the organisms that take the nitrite to nitrate. So, where you have is a very close cooperation between two different types of microorganisms, and the transfer of one of the substrates that's a product of the metabolism of one of the organisms to another one: so extremely efficient process that really is very important to take into consideration when you want to understand processes like sewer treatment, but also nitrogen biogeochemistry and the environment. Any questions? OK, so for the remainder of the lecture I want to talk about genomics, -- and then in particular also its application to questions of environmental microbiology and environmental science. So first, what I want to do is give you a little bit of the definition of genomics, and then cover how it is actually possible that we can sequence entire genomes, and I want to give you some highlights of what we have found by comparing different genomes to each other. And then I want to talk about this field about environmental genomics where we can use genomic techniques to actually learn something about the function of different uncultured microorganisms in the environment. So first, our definition, it's basically to interpret or to sequence, -- interpret, and compare whole genomes. And as you will see the comparison part actually plays an increasingly important role because we have now actually genome sequences available from almost all, or from at least some of the major groups of life. So this, again, is a different kind of representation of the tree of life. You have bacteria, archaea, and eukarya again. And as you can see, we have a lot of representatives. In fact, this doesn't even come close to the diversity that we have now sequenced as well over a hundred bacterial genome sequence now, several archeael genomes, and increasingly also in eukaryotic genomes. Now, genomes, so how is this done? How can we actually sequence genomes? Well, on the face of it we use very large facilities where you have sequencing machines present. There's one very important one at MIT, actually at the Broad Institute, and here you see all those really industrial scale production lines actually. But the basic problem is that genomes are large. E. coli, for example, has roughly 4.4 million base pairs, and the human genome is even much, much larger. It has about 3 billion base pairs. OK, so genomes are very, very large. But a single sequencing reaction-- -- gives you only roughly 500-1, 00 nucleotides or base pairs. So how is it that we can actually sequence entire genomes? I'm going to walk you through this, and there is some variation on the theme, but this is still a major approach that's still used in some of the sequencing facilities. Now, you start out by extracting genomic DNA from organisms, and then you use restriction enzymes to cut the DNA into relatively large pieces of DNA, so about 160 kilobase pairs long. On average, this is shown here. Kilo means a thousand, so 160,000 base pairs long. These pieces are then cloned into specific cloning vectors that are called BAC vectors. So therefore, cloning large pieces of DNA, and BAC stands for Bacterial Artificial Chromosome. And what they basically are, are plasmids, very special plasmids that can carry large pieces of genome, or large genome fragments. So, by cloning into those BAC vectors, what you do is you basically divide up the genome, and then the step number three is mostly done for eukaryotic genomes because they are so much larger. You can actually map and analyze the fragments, and map them onto genome maps where you know the location of different restriction fragments and different genes, actually. For bacteria, this step is mostly skipped, actually. What you do with each one of those BACs, is you cut them further up into 1 kilobase per fragment, so much smaller fragments. And these are called, and these are cloned then into normal plasmid vectors. And so you generate what are called shotgun clones. So, these are then cloned into E. coli, you go through the same type of steps that we discussed before already with environmental clone libraries. And you can actually determine the sequence of each one of those pieces of DNA. And what you will then get, is small fragments of overlapping DNA sequences. That it shown here. You'll find overlaps, basically, which piece together the whole genome. And so, first to assemble, you piece together these genome fragments that are present in the BACs, and then finally you piece together the entire genome propose large sequence pieces, and you get a so-called draft genome sequence. The next step in this analysis, then, is that you do so-called genome annotation is. And the first very important step is that you translate the gene sequences into amino acids. So, the nucleotide sequences into amino acids particularly in prokaryotes. This step can be done right away -- -- and what this allows you to do, is you can look for what we call open reading frames, or ORFs. And what you look for is a start codon and a stop codon that basically branches or frames a stretch of amino acids encoded by the nucleotides. So you look for ORFs. And these are your putative genes. The next step that you can do, then, is you can go to databases and now you compare your ORFs to information that is present in the databases. So basically, you inquire the database and ask, is a gene sequence that is similar to the one that I have statistically significantly similar present that allows me to say something about the function of this particular gene? So function, can then be identified by comparison with databases. Any questions? OK, so that allows you, then, to basically say something about the different genes that you have found in the genome, but to give you an impression of how new this field really is and how little we still know about the diversity of genes and organisms, on average when we sequence a new bacterial genome we find about 30% of the genes, or a third of the genes have no known functional analog of the databases. OK, so there's a lot to learn about the diversity of life and about the functional diversity of life. In eukaryotes, there are some little twists, as you all know. And basically, that is that genes of course consist of introns and exons, right? And so it's basically

relatively difficult to directly identify those open reading frames. And what you have to do is that you have to actually oftentimes, so let's write this down. And what people oftentimes do, then, is that they search for matching sequences in so-called cDNA libraries. Now what are cDNA libraries? Let me just show you this on the next slide. Skip this. Basically what you can do is you can isolate messenger RNA from cells and that translate the messenger RNA by a process called reverse transcription that the viral enzyme that translates RNA into DNA, so you can translate it into DNA fragments. And you can then clone those DNA fragments into plasmids, sequence those, and then basically see what are the pieces that are actually, what are the introns in the genes? What are the pieces that are excised when the messenger RNA is actually created from the genome? And so, let me just cover now a few of the major insights that people have come up with. Of course, it's a very growing field and a lot of excitement is coming out. And I first want to talk about bacteria and archaea -- -- and then say a few words also about eukaryotes or eukaryote. First of all, what we learned about, bacteria and archaea, is that their genomes are very compact. Whenever they have pieces of DNA that are not frequently used, they're actually lost from the genome. OK, so they lose genes, I should say, relatively easily, and we can see this that the genome size is correlated to metabolic diversity. So, for example, we have Mycoplasma genetalium and Streptomyces -- coelicor are two very different bacteria. The first one is an obligate intracellular parasite. OK, so, which means it's actually bathed in a nutrient solution in the eukaryotic cells that it invades. It doesn't have to make amino acids. It gets it just from the host cell. And it turns out it has a very small genome, so only 0. 8-based mega-base pairs, so 580, 00 base pairs, and only 517 genes. And interestingly, actually people are now using this organism to try and ask, well, what's the minimum number of genes that organism can actually will live with? And so, they are deleting in a stepwise fashion the different genes in this organism, and it turns out that you need about two to 300 genes minimum in order to make the things survive. On the other hand, streptomyces is a soil bacterium -- -- has a very complex lifestyle, can degrade a lot of environmental substrates, and it has a very big genome, one of the biggest bacterial genomes. And so, those two organisms basically span pretty much the range of bacterial genome sizes. And so, it's thought that it has about 7,846 genes. Now, we also have a very large genetic diversity -- -- between species. And typically what you find is that roughly 15 to 30% of genes are unique to a specific species. And that's really because bacteria and archaea have the capability to affect a lot of chemical reactions that eukaryotes, for example, cannot. There's about 20 million known organic substances, organic chemicals, and almost all of them are biodegradable by bacteria. Even the minutest compounds if it were not biodegradable bacteria, would build up in the environment, OK? So, if it just where a cofactor that some organism produces because we have such a long period of time of evolution on this planet and evolutionary history, you probably would be able to dig it up in your backyard. One of the other very important and interesting insights that has come out with comparing genomes for microorganisms is that lateral gene transfer is a very important process amongst microorganisms. Now what do I mean by lateral gene transfer? It basically means that we find evidence among bacterial genomes that they have actually taken genes from completely unrelated organisms. And I just want to show you one example here from that of thermotoga maritima -- -- which lives in hot springs. This is a very interesting bacterium that lives in hot water of around 80âˆžC and thrives only in those kinds of environments. And they coexist there with many archaea. And when people sequenced the genome of thermotoga maritima what they found was that about 25% of the genes have their closest relatives in archaeal genomes. So roughly 25% of genes in thermotoga are of archaeal origin. And how can we actually figure something like that out? Well, the most important technique is, again, phylogenetic tree construction. And so when you have, for example, gene A, well let me draw this, actually, on a new board. So you're comparing, say, three organisms, organism A, B, and C and you compare gene one with gene two. And you notice that most genes adhere to this pattern, but that every now and then there's a gene that gives you this type of pattern. What you can then conclude is that this gene, C, has not coevolved with the other genes in the genome of these organisms but was actually transferred into it from another source. And I don't have time to go actually into the mechanisms. If you're interested, I teach a graduate class that undergraduates actually take in our department, environmental microbiology, where we discussed a lot of the mechanisms. It's basically a lot of viruses can affect gene transfer but also plasmids and transposons. But for bacteria, again, you should remember that often new function is actually oftentimes arises by lateral gene transfer. And one of the interesting things is that lateral gene transfer is actually very important in the evolution of pathogenic bacteria. So, the so-called virulence genes, which are the genes that basically affect pathogenesis. Do you have a question? Among pathogenic bacteria, often arise by lateral gene transfer. OK. Any questions? OK, now for eukarya, I just want to make the point that their genomes are generally orders of magnitudes larger -- OK, and that the exons, so the stretches that really encode the protein that make up the organism, the exons are only typically a few percent of the genome. That's particularly in higher eukaryotes. Yeasts, for example, have a much more compact genome also. We, for example, are full of DNA that people still have a very hard time figuring out what that actually does. But it seems that the majority of the genome, so-called repeated sequences -- -- many of which seems to be ancient retroviruses that have inserted themselves into the genome and have since then lost actually their function. OK, so the remaining time I want to just give you an example of how we can now use these techniques that I outlined before to learn something about microorganisms in the environment. It's called environmental. Basically, the way this all started was by going into the environment and extracting nuclear gases and treating them exactly the same way as if you had a single genome. But, again, remember, we have a very large mixture of microorganisms present in the environment. And where this is mostly done was in the ocean, actually. And what people did, was they constructed those BAC clones directly from the environment and then looked amongst those BAC clones for specific 16S ribosomal RNA genes. Remember, this is the marker that we have for microorganisms in the environment. We know the diversity of microorganisms through those types of genes, and we have a lot of the data available. And so, in order to link a specific function of such an organism that we only know from the 16S ribosomal RNA genes. So, to ask the question of what much of this organism might be carrying out in the environment, it's very useful to sequence BAC clones that have 16S ribosomal RNA genes on them, and

environment, it's very useful to sequence BAC clones that have rRNA genes on them, and determine what kinds of protein coding genes are on there that might reveal some of the function of the organism in the environment. And one example that I want to show you is that of the proteorhodopsin. So basically, the initial task was to sequence BAC clones containing ribosomal RNA genes, and look for other genes that might reveal some of the function. So, you don't want to look for all the genes that encode proteins that are important to the cell cycle and things like that, but really sort of metabolic genes that might tell you something about the type of metabolism that this organism carries out in the environment. And so, what the first example that turned out to be really, really important is that people found rhodopsin genes on one of those BAC fragments, and it turns out this rhodopsin catalyzes or these rhodopsin genes produce a protein that inserts itself into the bacterial membrane, and it's a photoreceptor that when it's hit by light, it actually becomes a proton pump. So, it expels protons from the cell interior to the outside, and you already know that this is important in energy generation in all living cells. So proton gradient across membranes basically give the cells sort of a battery status that can be exploited by ATPase molecules or ATPase proteins that equalize the proton gradient and affect ATP synthesis in doing so. Now, why is this so important? Well, it turned out that this type of protein is present in almost all microbial cells that were previously thought to be heterotrophs alone in the ocean in the parts of the ocean that receive enough life. And what this means is that our estimates of the global carbon budget of the ocean were basically wrong because most microorganisms in the ocean have this. So most prokaryotes in the ocean have a light-driven proton pump which is called proteorhodopsin. And it basically allows them to gain energy from sunlight. And there's an increasing number of such examples now where we are learning to interpret environmental communities, and the function of environmental microbial communities through those genomic approaches. And it reveals basically an enormous diversity of organisms out there. And what we also are learning to do now is to assemble entire genomes from those samples by applying genomic techniques. And this is an example here where you see, this was published last year, where people went out and basically were able to piece together from pieces of genes obtained from the environment, entire genomes or fragments of entire genomes. And that's shown here. Those are contiguous sequences. OK, so if you have any questions let me know by e-mail, or if you're interested in pursuing this further I also teach another class in civil and environmental engineering.