Plymouth 9/04/2015

I: Interviewer (Gregor Halfmann)

R: Respondent (Rowena Stern)

I: I would like to talk about DNA barcoding a lot and I am not very familiar with it, so I would be happy if you can explain a lot of details to me.

R: Alright. Of course.

I: But first I would like to know how long you have been doing DNA barcoding and how did you get into it?

R: I have always been a geneticist but I applied my skills initially to veterinary parasites and then to human diseases, but then due to various circumstances, I was coming to the end of a post-doc and I was looking for another one, that was when there was a big project on DNA barcoding and this was for all sorts of life forms, mostly butterflies, but a lot of microorganisms. I read more about the system I was looking at, these types of algae that have very interesting genetic systems and I decided that I really like this area. I always had a microbiological interest from my PhD, so I took that job. I went to visit the PI there and that was when I started the project. I started that project in 2006 so I have been doing barcoding for nearly nine years or just over nine years.

I: When was the method of barcoding established?

R: Well, technically, I guess the paper was written in 2003. I think people have earlier than that been looking at genetic identification, although with a slightly different emphasis, since the early 90s or even earlier than that. It is just a slightly different application. What people used to do was genetics to confirm some sort of evolutionary information about the identity or the evolutionary relationship between an identified organism and its nearest relative and how that fitted in with the whole family or a particular taxa group. Barcoding is slightly different than that. In those cases, people used whatever part of the genome they liked and whatever was useful for their particular taxa and then did the research. This one was more organised, so it is using the same part of the genome and it’s a small section of that genome that a group of scientists agreed on based on an initial research effort. They said “okay, we are going to use this part of the genome to look at all organisms.” And the power of barcoding is that it uses this short section of a genome as a proxy for species identification. So it is saying “this mark can be used to say whether this species is x or is it not.” And the power of it is the database that is associated with DNA barcoding. The idea is to collect as many what we call voucher specimens, known examples of whatever taxa we are looking at, and create many and many of them. And just statistically being able to get your unknown organism compared to this huge database is the power of DNA barcoding. So it really answers the question “Is this species x or is it not?” What it does not answer are evolutionary questions: “How is species x related to this family?” It does not work well for the more sort of deeper questions. It is simply an identification tool and it does not work well if you don’t know what it is and where you can fit it in. It is also a way to make this accessible to non-specialists. You could essentially, if you are a micro-biologist maybe you could look at shark DNA and use the same method and be able to tell if this is species x or not, in any system. That was the idea, but it was a bit more complicated than that.

I: The next question is from a rather naïve point of view, I would say. It can be used to ask questions of biodiversity, right?

R: Hmm.

I: And how would you chose, when you take one species and then you go to the next species … How do you actually proceed? What considerations go into what you are doing next?

R: Yeah, I know what you mean. How does the pipeline work?

I: Yeah, I wonder … There is probably no end to barcoding, right?

R: No, when do you stop? When do you know what is enough? That’s a really good question. It is not naïve at all and people don’t know it. It is very much based on your initial assessment, but that is based on practical considerations of accessibility to the organisms. Say, you are looking at an organism x and the first thing you have to do is assess … because not all organisms are accessible to barcoding, they have particular genetic characteristics that make them sort of not ideal for DNA barcoding. So you assess a range of known organisms and for example, in the case of traditional barcoding this gene COX1 is used, “is COX1 a good marker for identifying my taxa?” So there are recommendations, for example, you should have between five and ten samples of any one species to look at basically for its variability. Sometime you find that one species is actually two species. And it is a little bit more complicated, because in any genome of any taxa, some parts of the genome evolve faster than others. So you might have a gene that artificially shows that particular taxa belong to two species, but it is actually two populations within a species. So you basically try to get as many examples as possible. Maybe five years ago you would use a sequencing method and you would base it on the expenses considerations and the effort considerations. You would say “maybe two hundred is enough” or “I would look at five or ten of my species,” but most people want to look at many species within a genus, so they say “okay, maybe we will look at two hundred” and they do all this sequencing and they ask “does this method work for my species?” and then they can test a series of unknowns. But now we have this new next-generation sequencing and the bar is higher now. People are now looking at thousands or even millions of sequences, although that is it. The older method was more labour-intensive, so you would have less samples, but each result was more accurate; whereas now, you would get more, but each result is less accurate and you have to use quite complex statistics. I don’t know, have I answered your question?

I: Yes, absolutely. You have already mentioned a few practical considerations. That was my next question. Can you guide me through the practical process of DNA barcoding? What is the sequence of necessary actions and practices?

R: Yeah, I think I will show you a diagram, too, because I think that might help. I recently had to give a talk to non-geneticists on how to apply genetics for species identification. … Sorry, it’s taking a while. So this is one of the things you might need: You start with acquiring a sample, let’s just say it is fish. Traditionally for barcoding, you actually have to taxonomically identify it and create a voucher specimen. So there is a particular set of rules, you have to describe it in a certain way and you describe certain characteristics. Then if you have got a big organism, you can get a piece of it and normally … I think the zoological code and the botanical code require a DNA barcode or a molecular signature to go along with the taxonomic description. So you get your piece of DNA and you extract it. There is a centre for DNA barcoding and it is in Canada in Guelph and they have given protocols. But in reality, I think everybody uses their own protocols because they have particular … In the Canadian Centre for DNA Barcoding, they focus on certain organisms and other people are finding that different methods are working for their organisms. So they do a DNA extraction and that is this section here and this is where we are talking about the PCR amplification. This is what is called DNA identification bit. They are what we call primers. These are little, little markers of DNA, small sections, small pieces of DNA that select the barcode region of the genome in any taxa. So they sit on a certain part and they amplify it, so basically, these reagents create copies of a certain part of the genome and create enough of it to be identified using what we call DNA agarose gel electrophoresis. So you need multiple copies, you cannot just work with the copies and the genome, you won’t be able to visualise it. So it basically creates more of the barcoding region. You run it out on a gel and you see and basically you can identify your section of DNA by its size. So if you say, for example, the barcoding region is about six hundred nucleotides long and if you run it out next to a series of known nucleotide markers of a certain size, then you can say “okay, I have amplified a piece of DNA that is six hundred nucleotides long, so it is probably my DNA barcode, but I don’t know for sure because I only got a size estimation.” So then what they do is DNA sequencing. Actually, you send off that six hundred nucleotide piece of DNA to a company or you do it yourself and it shows you the nucleotide composition of that particular piece of DNA. So from that point, then you have to analyse it. That’s the next stage. So, this isn’t a great resolution here, but what you do is … For example, the Barcode of Life Database is quite well organised. So you can use several ways, but what you can do is, you might set up a project. It is basically a software tool where you deposit all your sequence data and then you can compare it to your own specimens or you can compare it to other specimens in other projects depending on whether you are allowed to. There is also a publicly accessible comparative database called GenBank. So you literally copy your sequence in and it compares it to everything in the database and it gives you a score. It says “okay, your DNA is 96% similar to, let’s say, a salmon species.” Then that is pretty good. But at that level you kind of know “ok, I didn’t sequence contamination, I didn’t accidentally sequence a [unclear], I have got the thing that I wanted to get”. But then you want to say “okay, the next question is, is it salmon species x, salmon species y?” So what you have to do is a pairwise comparison. It is basically a chart of … I wonder if I have got it here, actually …

I: Would you mind sending me this?

R: I certainly will. So for example, here are all your species and they are slightly different to each other and you want to know whether a, b, c, d are the same species or are they different. So it is somewhat of a relative comparison because it all depends on whether you have got a, b, c, d or whether you have got two hundred more sequences. So you get what we call a genetic distance chart. If you compare a and b, you can see they are identical. And you say “okay, they are the same species.” If you compare a and c, there is some difference, there is an a here, there is a t here, there is an a here, there is a c here; so you can see, there are like three differences between a and c, and b and c also are three differences, and so on and so forth. But the point is, you have got all your differences, but you say “where is the cut-off? Where are two species identical and which ones are not?” Luckily, for many species, particularly animal species, people have already done that research and found out that there is a general cut-off rule. So they said “species with one percent difference to each other, they are probably the same species”. And what they found out is that there is a big gap between members or individuals within a species and individuals between species. So members of two species are going to be much more different than members of the same species. So members of two species will be ten percent different and so you can do a cut-off. You can say “I cut off at one percent difference, anything more than that is another species”. So for example here, in this case if we use the animal criteria, a and b can be the same species. Obviously, this is a percent, so it all depends on how many … I have just used a short piece, but maybe a and d are different species. Depending on what your cut-off is, say two percent is the cut-off, so a and b are the same species, maybe e and d are the same species here, but sometimes you get a bit of overlap and you are not sure whether it is one species or another. So that is the point where you get ambiguity. So there is an assessment of whether … You got your sequence identity, but then how good is it at discriminating the species and it is all statistical. So you can say “that’s the difference within a species and that’s between species.” And there should be a gap, there should be nothing in between here. If you are getting overlap, you don’t know whether these things, species that show intermediate kind of genetic distances between each other, belong to one species or two. So that’s the issue … I might get some water so I’ll be back in a second.

I: Okay.

R: Stop me if I am confusing you. So supposing you have established your example, you have got a good system for identifying the differences within a species and between species. You simply apply a cut-off and say “okay, I am comparing everything of fifteen individuals and if they are within one percent difference, they belong to salmon x.” So that’s the process. I will say there are a lot of, how could I say, inter-individual differences. And a lot of people are confused by the difference between what we call molecular phylogeny and DNA barcoding. Molecular phylogeny is a way of, kind of using similar methods, but it uses them to answer different questions like “what is the evolutionary relationship between species x and y?” rather than “using this DNA piece, is this species x or not?”

I: We already talked a bit about the part of analysis. I would like to go back to the practical part. You showed me the different steps in the beginning. What is the temporal dimension of these steps? How long does this take?

R: Right. It depends on how easy it is to analyse your sample. So, let’s say everything is working well, it may take you, depending what system … So for me, I work on microbes and I’ll just describe what I do. So I go out to the sea, it may take a day, and collect literally just a jar of water. That is enough for me. I concentrate it up and look under the microscope just to generally see what is there. And then it might take a couple of days for me to pick out individual organisms. It depends … So that would take me two days, it may take me a week then to do the DNA extraction and another week to do the sequencing. So that may be two weeks in total. And the analysis is a bit longer. It depends on how many individuals you have. I could do that for maybe twenty individuals, but if I was, say, working on sort of a scale factor of hundreds, it might take me, if everything is working well, two months to collect all the data. And it will probably take me another two to four months to do the analysis. So there is always a bit of a lag. You are collecting and by the time you get your results, sometimes you are in a position, where you are like “okay, well that didn’t work, but then I don’t have enough time in my project to go back and reassess.”

I: You said if everything goes well, but what are the things that can go wrong?

R: Alright so, you might find yourself working with a set of organisms, which are very difficult to analyse, for one reason or another. For example, jellyfish are difficult to analyse for two reasons, because they are made of a lot of water, somehow it makes DNA extractions very difficult. You can get DNA but it might have a lot of what we call inhibitors, so some chemicals that stop the PCI amplification reaction and you need a PCR product in order to get your barcode. It might have like jellyfish, they are kind of ancient organisms, and then that particular marker COX1 does not work for them. So those people who did that project probably had to go back and look for another marker that does work, so that would double the time. And they were probably looking for two or three markers at the same time. So it might take two years. When I was doing my barcoding project, my boss said to me “this is an interesting barcoding project, but when do you know when to stop?” And we stopped at the point when we knew that there was a competitor who was writing it up. So we had to stop it there. But we stopped when we collected all the samples we could collect within reason and then basically based on the project deadline. So you have a three year postdoc, you have certain objectives to finish certain points. It took me about a year or a year and a half to collect and sequence them all and then it actually took me much longer … So when I say four months, that’s because I am used to DNA barcoding now. When I first started it took me a long time, because I had to evaluate the barcode and then there were certain circular arguments in there. So a lot of people gave me samples and said “okay, can you analyse this? We don’t know exactly what we think. Is it this species? Can you use the DNA barcode to answer that?” And that’s the wrong way around. For me it is to know what it is in order to say “okay, confirmed this is [unclear].” But sometimes it is the identification: So what can also go wrong is your taxonomy if your original … DNA barcoding is based on taxonomical identification first and if your taxonomy is wrong then DNA barcoding can be really confusing. You have to use some sort of reference to identify it. It cannot be used on its own for identification.

I: You partly already talked about this. I want to talk again about the different steps. Which ones of these are you actually doing yourself or mostly yourself? You said you were going to sea by yourself and taking a sample …

R: Yeah.

I: And where do you rely on the work of others?

R: Right. So often you have to rely on people giving you samples. So that is the main point where somebody else is involved, sample acquisition. Probably because you can’t physically go everywhere in the world and collect everything, so you rely on people who have what they call culture collections. If I was working on animals, they would be museums or something. And you are relying on their identification and you are relying on the fact that they have maintained the organisms correctly all the time. A lot of them told me they were quite worried about it. It was not always the case that they were maintained, because there was a lot of funding issues, a lot of changes in staff and that meant that identifications were often confused or maybe contaminated; one algae culture was contaminated with another by mistake. So I relied mostly in my project on other people for giving me samples, but everything else I did, right from the analysis, from doing the DNA extraction, to the analysis, to writing the paper.

I: So you also already partly answered the next question. I was going to ask about the different sources of samples that you use. What are the characteristics of each or the constraints? You already mentioned museum samples, for example …

R: Right. So, for my things … DNA barcoding is not so straight forward when you are working with microbes. One of the things that is missing and that you don’t have … You are dealing with a single cell. So when you have done analysis on that cell, you have used it all up and you can’t say “I have got a similar one and I put it in the museum.” And a lot of cells cannot be cultured, they cannot be kind of preserved in a way that you could look at it later. So there is that gap, but basically, you do not have a voucher. You just have maybe a photograph which is not the same thing as a living thing, but it is the next best thing, and then you have your sample. So I only get it from culture collections or I get it myself from my local environment. You know, sometimes when I go on holiday, I collect some. So it is somewhat random. Basically, it is based on resources rather than the scientific question. Most of my samples were from northern latitudes, simply because that’s where most of the research effort was. That is not to say that that is where most of the diversity was; that is probably in the tropics, but … Where else? I got it from private researchers, too, who had their own cultures and said “oh, we are quite interested in your work.” And I met people at conferences, other scientists, and they were working with museum collections like the Smithsonian. This lady had a corral collection and I worked on algae that lived in the corrals, so she grew these corrals up with the algae and gave me some of those.

I: So there is this notion that biology or ecology have become so computer-intensive that the role of going out and getting your hands dirty is kind of diminishing. You do go out yourself, but what is your opinion on that?

R: It very much depends on what you are working on. I would say that probably for most scientists, even those working on large organisms that require a lot of [unclear], most scientists look like a bank manager. They are doing computer work all the time. So in my job right now, in fact, I work for an organisation that has specifically the task to collect plankton using an autonomous system. So I don’t go out at all, I don’t need to go. The only reason why I might go out, it is occasionally, maybe once or twice a year, because the samples that we collect are dead, is to collect live things to keep them alive in cultures. So there is less of a need for that. There are still people, who go on cruises, but again those are … They are either one-off … I am writing a paper on this, because it is an interesting question of what motivates us and what kinds of reasons we use to go out. You know, some people go on cruises, but again, that is like maybe once a year. They might collect a lot of samples and then they do a lot of analysis. Maybe they go once a year and then they won’t go on another cruise the next year. It’s like once every two years. So yeah, I would agree that going out there and doing sampling has become less and less.

I: I was going to ask more about SAHFOS in particular and what it is like to be here, how it is actually working. You mentioned that samples are collected automatically, basically. So are you being asked then to analyse these samples or do you pick whatever you need for your research questions or is there a rate of samples coming in?

R: Right. So what is the process and what are the selection criteria? I would say there is a bit of both. So because the samples are … A lot of other institutions are very much hypothesis-driven, by questions. So they are collecting in a way that is designed for a particular project. So often that leads to a series of samples that are taken. Most of the samples are taken in the same way, because there is some sort of general agreement, but often there is not any consistency between samples. SAHFOS is all about consistency. So we have monitored in the same way for eighty years. We have collected using the same device, using the same type of preservative, and have a lot of quality assurance processes to make sure that the analysts, and that’s an area of variability, look at the samples and analyse the samples in the same way and that they are getting the same species. They even have … If there is a person, who has analysed a sample and the results are looking unusual, they will go back and they call it check block. They will go back and a more experienced analyst will look at this and see if it was wrong and they will go over it again. So I basically use that resource for certain research questions, so I’ll pick some of them. I won’t do it consistently and take, whatever, a hundred samples and do DNA extractions for whatever use. Instead I would say “I am interested in this particular species at this particular time period.” So I am going to do DNA extractions of these certain samples. But there is another autonomous sampling system that goes along with the CPR and we call it a water micro-plankton sampler and that goes on a certain route. That is somewhat regular. Every month it comes in and I analyse it in the same way. I think it is very important for a long-term time series to analyse things in the same way because if we went by research questions all the time, say in the 1880s, our idea of what is important was very different from what it is now and now doubt it will be very different a hundred years hence. And if you keep changing according to what is the latest idea, you won’t get any good answers.

I: I was going to ask you about the water sampler later. Now more about SAHFOS, this is a bit hypothetical. Could you do the same work or similar work, if you were not at this institution? Or in other words, what is the benefit of being here in direct contact with those who collect the data?

R: Yeah, that’s the key question, I think. That’s the point. I could do my work anywhere, but I couldn’t do it … Well, I could. We do have an associate researchers’ scheme. Somebody could come and people have come, for example there is an Italian scientist he used to work here, he is aware of this collection, and he says “okay, I want to look at all these samples for my particular question.” So I could do that but probably, in reality, not many people are aware of this collection. It is somewhat like a museum collection. So there is probably a limit. If I was working outside, they probably would not give me everything that I wanted. Here I have a bit more freedom to analyse what I like. But also, I get updates from the analysts who tell me “oh, I am getting a lot of fungi on these samples” or “I am getting a lot of this on that” and they sometimes come to me and say “this is an unusual organism.” So there is some sort of synergy there that you won’t get elsewhere. And certainly, the questions that I am always interested in, like long-term changes over multiple decades, I probably would not be able to get, or only very rarely get that at any other institution, except for perhaps the Bermuda Atlantic time series or the Hawaiian time series. Those are probably the two other organisations where I could do what I do.

I: Does the fact that SAHFOS is a charity affect your work in any way?

R: No. To me, the charity is just simply a name. For me it is a research institute.

I: Okay, fine. So going back to the barcoding again … If I understood correctly from one of your publications or one that you co-authored, you were part of a working group that wanted to develop standards for barcoding. I am wondering why that was actually necessary, what were the difficulties and also, whether that was successful or not?

R: Very good questions. In the area I am working, microbial ecology … and when I say microbial I mean that there are particular types of microbes which we call eukaryotic microbes. That is basically stuff that is not bacteria or viruses but are still sort of single cells, that’s what we are interested in. Not a lot of people have done research on that and also, the organisms … I can show you maybe another presentation … These organisms are highly diverse and come from multiple kingdoms. … Let me see … Alright. So if you look at what most people barcode, it is this, metazoan, and the group of organisms I am interested in has a huge diversity, but they are all mostly microbes. So people were coming together and they were saying “well, cytochrome oxidase is the standard barcoding gene that we are supposed to use, but it is not working for my organism.” And there were a lot of people who were finding this. Also, the barcoding organisation came to us and said “actually, you are supposed to be doing this. This is the way.” But we could not do it for various practical reasons, because as I have told you before, when you do a barcode you are supposed to officially collect it and preserve it in a certain way and describe it and put a physical specimen in a museum as the voucher. Then you are supposed to do the molecular identification, describe it in certain ways. What we were finding was that the organisms were not even … we didn’t even know whether they were part of the zoological code or the botanical code because those were set up in the early, I don’t know, early 1900s. So actually, there were a lot of things that were kind of intermediate. The issue was basically that there was a taxonomic confusion within the group of people that I worked with. People were working on similar organisms to me but how do we deal with the fact that no one knows what they are working on really? Out practical considerations were how do we create museum and voucher specimens when we are looking at single cells and we can’t physically give a voucher? There was a person who was from barcoding, I guess he worked for something the Smithsonian Museum was part of, and he said “none of you are doing barcoding correctly because you don’t have a physical voucher.” Then we agreed, “okay, we need to have a photograph instead and that’s the best we can do and then we also need to make sure we do much more joint analysis between the molecular and the morphological.” Because a lot of people only do genetics or they only did the morphological. There wasn’t a thing to do both, but there was also the question you asked: “How many do we use? For our group which is a heterogeneous group of organisms belonging to multiple large kingdoms, what marker do we use?” In the end it was not ideal. A lot of people disagreed but the majority said “okay, we are going to have to use two barcodes. We have to have something standard and we are going to use this gene” which is, I think, a part of a ribosomal, a small sub-unit ribosomal gene and a particularly variable section of it. That’s the primary marker that everybody has to use. So at least there is a comparison alongside. And we picked that based on the fact that there was already a large database there. Then we said “okay, everybody else is going to have to use their own specific DNA barcode marker for their taxon group.” So that was the part … When people do that, that means they cannot cross-compare to other groups, but it will at least give you some accurate identification, whereas the other marker will give you a cross-comparison, but not so accurate identification. So we had to compromise.

I: You mentioned in an email a few months ago that you were also in discussions about the usefulness of the barcoding. Perhaps you already covered that a bit.

R: Yeah.

I: Or can you recall exactly what you meant? What was the discussion about the usefulness?

R: I think what it was … There are two aspects. What I just told you about now was “what is the use of barcoding for organisms like this? If we can’t have a barcode that does not identify our organisms well, why is it useful for us and how can we contribute to the DNA barcoding community?” We said “we could,” but we had to admit that the universal idea of one marker that can identify all organisms on earth is not possible because of the diversity of life and how different organisms use their genomes in different ways. We had to admit that there was some level of disagreement. But it is still very useful because in actual reality, most people won’t compare a microbe with a larger land animal like a big cat. They won’t do that, so in reality it is useful within a broad remit. In any case, it is just generally useful to have the same set of ideas, that people know the rules when they are generating what you call a metadata set. Datasets that scientists in different parts of the world should know. That way we can start to fill in the gaps in the databases. So maybe we have organisms that are described well in a certain way and then the genetic data has to be compliant, so that you can start comparing datasets and so that you can drill down to … So that was the usefulness of it. What happened at the end of that was that people wrote a paper, so a lot of people were interested in just getting a publication but not necessarily doing anything more with it and that is probably because a lot of people there were PhD students or post-docs. There were few group leaders but from their perspective, nobody knew whether they were going to work in the field for the next three years or not. In a lot of cases you find a project and then it stops after the end of, you know, when the interest wanes or when somebody leaves or the end of a funding comes. You get a lot of these projects that are kind of static. One other thing the Canadian Barcode of Life did well is, with Paul Hebert, just continuing on and keeping that thing going. It is an active database, it is curated actively and revised all the time. But that said, a lot of people don't use it, because they made a lot of tools for scientists, but those scientists are often not that familiar with the methodologies. They might just be new to this area and so they often find it intimidating and in the end they don't use it. It is there and it is useful for the people who know it, but it the barcode was supposed to be accessible and it has ended up being kind of restrictive just through expertise.

I: We already started talking about what I wanted to ask now about the databases. I saw the Barcode of Life Connect website and I saw that you have a profile there. Can you give me an idea of what that is? It looks a bit like a social network.

R: It does, doesn't it? Let's see if I can sign in to that. I have to say I used it when I was a postdoc, I don't really use it now, although I do get emails to say “oh, we have looked at your database and we see that this latitude and longitude coordinates of this sample are wrong, so we have curated it. Please check it.” So it is being curated, but …

I: So it is also a database?

R: It is supposed to be mainly a database. … I just log in. … So you can see these are all people I have shared with. I don't know these people but they have checked my projects and they have changed it. So these are all the projects that I did. I looked at maybe four or five groups of microorganisms. These are haptophytes, these are dinoflagellates, but I also put a dino trash; it would not actually allow me to throw away anything, so I had to create another folder called trash. This was a group of algae and these were also dinoflagellates. That one, that one, that one, they were the main organisms I was working with. On that one I was using the COX1 marker and on that one I was using a different type of marker. And then it is cryptified here. If you just go to one of them, it gives you kind of a breakdown of all your organisms, so it is user-defined. I put in all the information, but then I did have to provide external information which could be checked by the data managers of this database. So you can see, this is a taxonomic breakdown of whatever types of taxonomic orders and then it gives you some statistics. You can see project growth, you see I have not done anything for so long, you see how old they are.

I: So what does an actual piece of data look like?

R I'll show you, but do you see what I mean? This is not great. You would not necessarily know how to navigate this. They do have conferences, people who do barcoding probably work in genetics labs but it takes some time getting used to it. This here has changed, it used to be a lot simpler. So this is what a record looks like. These are all the things I had to put in … I wonder what it says here. Flags, so every record flag here is probably saying that it is not a proper barcode because I don't have any pictures in this case, because they all came from algae culture collections who refused to give me pictures. So what you do is you have to enter meta information associated with a sample. So this is where it was collected, this is where the culture collection is storing the sample and then this process ID is basically the identification just for this database and that is, I guess, the identification of the culture collection. So it gives you all these things and again, I did not have access to all information that they wanted. And this is the actual data, what it looks like. So that's it. It is literally a series of letters. That's a DNA sequence and then this tells you ambiguous characters which is basically the quality. If you have more ambiguous characters the worse it is. And it gives you the predicted protein or amino acid sequence which would be translated from a DNA sequence. And it gives you an illustrated type of barcode, I never used that; I don't know, I mean, it is just some sort of thing, but yeah, you can see here, I have got what you call a GenBank accession number. So if you go to GenBank … PubMed hosts GenBank, this is like a multi-database thing, it is a storage and database for America. It might even be directly linked, but I'll just see … So if you type in the accession number, okay, that is actually the paper linked to that but that is because it is on PubMed. If I go to nucleotide which has the actual DNA sequences, that is the raw data. That's the same information but written according to the style of this database. GenBank is by the way by far the most common and widely used database among geneticists. It is a bit more clunky, it does not have quite the pictures and the nice kind of things, but … there it is.

I: Yeah, I am wondering, since we are very interested in re-use of data. Are these data being accessed by other researchers to be used? And what makes them re-usable and for what?

R: That's a good question actually. I'd suggest that you actually ask the BoLD system directly, because they actually have statistics on how many people access at least their database. The only way I know about this is through … I am going to go back to the paper, it will show how many times it is cited.

I: Yeah.

R: But actually, I have no idea whether people use it or not. I am assuming, generally speaking, not. What people do is they like to use the general information from that project. Supposing somebody was looking at this particular organism that I showed … They might include that in their specific taxonomic study. That would be one way but less and less people are looking at these traditional taxonomic studies. A lot of people are using this next-generation sequencing to look at overall diversity in which case they are generating their own datasets. But they might look at my paper to see “Oh, should I use this marker? How did she do the analysis? I'll do it a different way ...”

I: Would you go to such a record from another researcher?

R: Yes, I would. I have actually looked at it. The record that is kind of useful is more the methodological development used for particular projects. It might be useful for another project. In the system here, you can see what databases … I am going to “workbench” here. They have a thing called “primer database”. This is really useful. For example, I am doing research on algae, but some people working here are interested in copepods; these are completely different. Let's type in copepod and see what I get. So from this I am seeing that the markers people are using are not COX1, but this thing called 12S or 28S; and F63, that is the actual primer sequence and that paper will have the methods, so I guess they will definitely use some metadata associated with that. I think people will reuse the data, but I don' think it is that common unless … Well, the type of research I am doing, maybe like twenty people in the world are doing it. So the number of people who are going to be using my specific raw data will be very low. But what might be interesting are the overall sort of conclusions.

I: So that probably must go with a lot of information about the methods. Does the database kind of encourage that? Providing that information?

R: It does. In fact, I have not done it properly. I gave in sequences and when you get a raw sequence it kind of looks like this. … Okay, so sometimes it is really nice like this and you look at this, it's fine. But sometimes you get bits like here … See, this is what a nice sequence looks like, it has nice peaks, but this thing is a bit messy and you might get a sectional sequence like that. So often you have to manually edit it to make it a bit better. The Barcode of Life, they like to see this, they want to see an edited sequence, they want to see every stage, which is really good. If a new user wants to drill down and say “okay, I am getting some confusing results. Did they get this wrong? Did they edit it correctly?”, they can do that. Whereas in GenBank, there is nothing, it is flat data …

I: I was going to ask how other databases actually do that …

R: This one is a bit of a problem. It is probably okay for certain species, like animal species, but for microbes it is probably not. It is very much user-defined. Somebody told me at that meeting that there is a lot of pressure in certain countries like China to publish. This might be an anecdotal thing and I don't know whether that is true or not, but he said he saw things that looked like they were deliberately mislabelled as something else. It was the organism Sellaphora, a diatom, but they were labelled as something else and he was wondering why that happened. He was suggesting that perhaps people were labelling things as something else possibly due to ignorance, possibly due to the fact that they needed to generate results and that was the result that they wanted. They say “we label it as that because my boss wants me to label it as that.” But generally speaking, people try to do it, 99% try to identify correctly but they don't necessarily know the ways to do it, so they will … I am trying to show you an example. Maybe it's in here actually … Here, you might get the wrong output. This is where I was saying you create your own database and then you have to see whether your marker works, okay? … Here, we have this example. That is the sequence. That is the GenBank database. This blast is a tool that can interrogate the GenBank database and give you the similarity. In this case, one species, it is Syllis vittata, a species of this group, and it is showing a hundred percent identity, but then going down a line we say “okay, this is another member of this family, another species of this taxa, showing less …”, but then we are seeing Lepidoptera. Lepidoptera is a butterfly, so the database is saying this kind of organism is similar to species of its own sort of family, but also to butterflies. And that is clearly a mistake. So a lot of people create a sequence, there is contamination and they mislabel it. That can be, but I would say that maybe depending on the organism you are looking at, for the mainstream looking at animals, looking at humans, it is probably correct.

I: My next question is perhaps a bit more general. I was going to ask what you consider as the actual product or the outcome of what you do. I mean obviously you are publishing papers, but how would you characterise what you put into the databases?

R: Yeah, that information … That is a very good question. I would not call it a product, no. For me, a product is a paper, it is a finished analysis of the project that I was starting on because I defined it and you get money from a certain organisation to do a project and the money is related to the project. So the outcomes are “I will produce a paper.” I guess for the barcoding project I did when I was in Canada, one of the products was producing barcodes, so I guess that would be a product. So I will say “yes, it is,” but in general, I would say, the raw data is not … it is part of … it's not …

I: It's not the end product.

R: Yeah, it is not the total end product. It is one part of it.

I: Okay, so what was I going to ask next? I saw your most recent publication with the water sampling device and I was going to ask about this technology. I looked at the methodology that is described in the paper and it looks like an incredibly complicated sequence of steps. It looks to me like a long series of decisions that you as a researcher need to make. I am wondering how many of these are actually determined by standards or where are you free to choose? Or is it an immediate result of what you found earlier?

R: Maybe I will show you this: I am part of a group and we are all interested in marine monitoring and there are a lot of people interested in molecular applications. They want to enhance the monitoring systems using molecular methods because molecular methods can identify a different set of organisms or many more organisms to a greater detail. So they want to enhance it because it might increase their chances for getting funding through, you know, novel publications. But essentially, no, there is very little to no standardisation. I gave a questionnaire to everyone, major researchers who use molecular biology, or barcoding, or what they call metagenetics which is, I guess, kind of a high-throughput barcoding method to analyse their examples. And everything was highly variable, except for perhaps the way that they preserve the samples and the pipeline that was used. So the analysis pipeline is highly technical and involves a lot of computer-technical knowledge. I think a lot of scientists are not trained in that, so they tend to use published methods. A particular published method that was easily accessible, that involved less of your own personal decision making, it led you through a stepwise protocol. A lot of people like that because it had some structure and they didn't have the expertise to do their own kind of analysis. But a lot of people then modified that for their own purposes. There was variation in even “what volumes of water do we collect?” It ranges literally from a hundred millilitres to hundreds of litres. There was variation in the geographic emphasis. It was not systematically. It was basically, because most research depends on project-based funding, it was researcher-focused. One particular researcher was interested in certain groups of organisms in the Arctic. There is a lot of research on the Arctic protists but hardly anything in other areas. So very inconsistent sort of decision making, there was no consistency in where to sample or how much to sample. I found it interesting what people did after they got their samples: A lot of people just sat on it for a year, then they kind of analysed it all at once, they produced a paper and then forgot about it. Whereas the issues that we brought up in that particular group, the ICES group, we are interested in long-term monitoring, so we are looking for extended datasets. We are also interested in using archive data, basically to re-use and to be able to get hind-cast analysis and to get information on past environments, so we can predict. There is very little of that. A lot of people analyse their samples and say “okay, project is finished.” And then there was very little effort in where you are keeping the samples, you know, are they curated? Nothing … I don't know, I got off on a tangent, but that's what I found. I might send you that paper, it is kind of a draft on, you know, not really any kind of general consistency.

I: I am jumping myself a little bit, too, so …

R: That's okay.

I: I was going to ask more about the water sampler which is if I understood correctly a new technology that is put on the Continuous Plankton Recorder. Have you been involved in any way in the development of that device?

R When I arrived here, it was there. They bought it and made the decision earlier. I think it was developed by CEFAS, a government organisation that carries out all the regulatory scientific tracking of organisms. They made a huge one which was supposed to collect water from a static station and I think [unclear] had a colleague and together they developed a smaller version. There was a question on whether you would collect the water on filters or shall we just collect a bag of water, raw. And in the end, they decided for the latter because they did not know what application it was going to be used for. This was before I came. So when I got it, it was like “we have this thing, can you use it?” So it is a slightly organic decision, I tried it out and I said “okay, it works,” so we started doing it every month. But no one explicitly said “well, this is going to be a new long-term monitoring tool.” We tried it out and it is still working so we just keep doing it.

I: So I am not sure if you have experience with the interplay between people who develop new technology for scientists to use or to collect data and those who then use it …

R: No, actually there is very little. Basically, you look at what is available. Some people make their own. A guy down the road is involved, he has a similar background as me, a molecular biologist, but he is developing his own systems; but he s sort of having to hire people, get students or companies to do it for him. But most people don't. In fact, just today I got an email from somebody in Southampton who for the first time ever said “we try it out, we write a grant to look at automated sampling systems and sensors. Would you like to get involved in a workshop? We would like to work with people to see what they want.” Most people look at what is available and you try to modify it yourself and use it.

I: You already partly covered what I was going to ask next. Again about the drivers that drive you towards a new project. You just mentioned it has been kind of a organic decision, so …

R: Yeah, well partly. I think there is …

I: I am also thinking about, we mentioned data availability, but also interest of funders …

R: Novelty.

I: Novelty, okay …

R: Yeah, there are scientific organisations and I like to see them as an entrepreneur or a business: You always have to have a unique selling point. And whereas a scientist, a molecular biologist not working at SAHFOS, might look at this or ideas as the unique selling point, [unclear] I guess, for us it is what type of data can we collect that is useful to other people and can generate a unique data series. SAHFOS is unique and is the longest time series in the world and that is probably what we are trying as our unique selling point. It is creating good quality and curated data for public use, but also for our use in looking at new connections or new ecological theories. Other drivers are, I'll be honest, there is competition. One of the drivers to get automated systems is that they thought SAHFOS was antiquated and old-fashioned and just that alone was a reason not to fund them. So we feel the pressure to be novel and generating data that is current and cutting edge. Because science funding is quite faddy. The new set of funding rules is now business orientated. So generating new technology is one of them. So we feel in order to be competitive and get money, in order to continue the survey, we have to be using new technology. But I guess it does benefit the survey, it might generate extra datasets and generate new research ideas from that. I guess those are the main reasons.

I: I will finish with some more broader questions, I don't know if it really makes sense to answer them. So thinking about the barcoding data, can you tell me about the impact that it has or that it already had on scientific knowledge, on what species are or what diversity is, for example?

R: Yeah, I like that question, I have to say that scientists are very much divorced from the impacts of their research. The end product is to produce a paper and being able to generate that paper in order to develop your ideas further. Using that paper and saying “well, I have done this research I would like to develop this idea further,” to kind of sustain your own research ideas, a lot of scientists are not interested in that. But I will say in the case of barcoding, because I was connected with other people's research in a broader community, it did impact the way other people designed their experiments. So a lot of people asked me afterwards “I want to look at these organisms, you found this, how should I do this now?” So for example, Tara Ocean Expeditions asked me beforehand; they were considering a marker but when I published the paper, they said “okay, we are not going to use this marker; we are going to use something else now.” In terms of species discovery, yes, you know, in terms of the actual science, I actually discovered that there were different aspects to the organisms that were unknown. I discovered that there was a lot more … There was a feeling, you know, that … people were kind of insecure, well not insecure, but … what's the word … maybe threatened by DNA barcoding as this quick and cheap method that anyone could use and therefore the area of science in general will be degraded. There would be no more need for taxonomists because you can just do this quick thing and you don't need any expertise. But I like to think it showed that there were a lot of question marks that the barcoding brought up about taxonomic questions and the fact that one outcome was that we need a lot more taxonomy because we don't know these organisms well enough. And I think it also brought together a lot of disparate culture collections, a lot of organisations that I got samples from, which could use that data. For example a culture collection curator came to me and said there was an unfortunate accident in the culture collection and a lot of their cultures died, but luckily they could use my data to say what was important and what were the key organisms to save. So I think in that way it was useful. But I don't know, I think I agree with you, it is all about who you connect to and how you involve your stakeholders to how the impact of your science … If you don't present data in the right way and if you don't involve them, they are not interested. You generate this data for other scientists, we are not interested in this aspect. Looking back, if I was to do this again, I would probably involve people from industry and policy-makers as well and say “how would you like this data presented? What is useful to you?” I think that is part of the reason why people don't use the data enough, they are not aware of it or they are not interested because they are confused about it.

(end of recording)