Genome Insider S4 Episode 6: The Megadata of Lake Mendota - Part 1: Many, Many Mers

Menaka: This is a production of the US Department of Energy Joint Genome Institute, the JGI.

Menaka: Right next to the University of Wisconsin, Madison. There's a Lake. Lake Mendota. It's big enough for boating. People fish and swim there. And this is Wisconsin. So this lake freezes in the winter and thaws in the spring. Just south of Lake Mendota, on its banks. A bunch of lake scientists have set up shop. This is the University of Wisconsin Center for Limnology. These scientists are happy to tell you --- Lake Mendota is the best studied lake in America, if not the world. And they're probably right. Since 1980, researchers have tracked environmental conditions here.

Trina McMahon: They study the fish, they study the plants, they study everything else --- but the microbes.

Menaka: That's where Trina McMahon comes in. She's a microbial ecologist at the University of Wisconsin, Madison and a JGI user. She wants to know about these microbes because they're the foundation of this freshwater ecosystem. And so there are lots of interesting questions to ask about them. How do those microbial populations change with seasons or temperature change? And what do those changes mean for the future of these ecosystems?

Trina McMahon: So I come along and say, "Hey, I'd like to study the microbes." And they say, "Great, just tag along." And so we have access to all their infrastructure, all their historical data. We have archives of samples going back to 1999, even for the microbes.

Menaka: This sample collection is like a whole reference archive of freshwater microbes. And this is actually Trina talking about this work a few years ago in 2017. At that point, she's talking about 18 years of microbial samples. A few years later, her lab would prep all their samples for sequencing and send them in to the JGI. Then the JGI would need to assemble this whole genomic reference archive, which turned out to be a giant task. After sequencing, this archive of information from Lake Mendota worked out to be 25 terabytes of data, a job too big for a single computer, or even 20 computers. At the time, it was the biggest project the JGI had ever assembled. So this episode is the first of a three-part series. The story behind a project as big as the one from Lake Mendota turns out to be too big for one episode.

Menaka: This is Genome Insider from the US Department of Energy Joint Genome Institute, where researchers discover the expertise encoded in our environment --- in the genomes of plants, fungi, bacteria, archaea, and environmental viruses to power a more sustainable future. I'm Menaka Wilhelm. And in the next few episodes we are taking a look at how this huge 25 terabyte data set from Lake Mendota comes together. Because remember, microbiologists sampled Lake Mendota for two decades to build this archive. So we'll get to the lake in our third episode of this three-part series. But there's another story here too. It's about overcoming the giant challenge of handling all of the data from those lake samples. This was so much information, it pushed the limits of the JGI i's most powerful programs. So stay with me. This little series is also a story of software and supercomputers. Because while biologists were out sampling Lake Mendota in their boats, data scientists and software developers at the JGI and Berkeley Lab were developing specific programs to handle this scale of data.

Menaka: They did that over time, not all at once. So this episode is the story of how those programs evolved to handle bigger and bigger data sets. And as we start, I have to mention the foundational fact that sits under all of this work. The software, the computing, the sampling. It's that sequencing has gotten better and better every year. That means sequencers hand back more and more data. And it also means that analyzing sequence data gets more complicated too. And here's why. To sequence a genome, you've gotta slice it up a bit. So the sequence you get back isn't complete. It's more of a raw material. To get a sequence sorted and stitched back together correctly. It takes software. So these programs work with data from sequencers and walk through the steps of sorting and stitching all of those sliced up sequences back together. These algorithms get way more complicated when there are more sequence slices. More pieces, more possibilities.

Menaka: And the goal is to get all of the slices into the most correct arrangement. So developing software like this is a really key thing that the JGI does for its users, both for this project at Lake Mendota and other genomics researchers too. There are entire teams of data scientists and software developers at the JGI that are constantly leveling up their programs to handle more and more raw material. They want to do that accurately and quickly enough, because more information could mean more understanding. And there's a real benefit to working with big data sets. Getting a zoomed out view of a dataset gives you a more complete picture of what's going on. It's kind of like if you have a photo and you can only look at one square, you might think, okay, this photo is the top of a black bear's face. But if you can look at more of the photo, you can see that the black bear is sitting in a dandelion patch.

Menaka: How cute. This is especially true for environmental genomics where you sample an ecosystem like Lake Mendota. There you really want to look at the whole picture. And also you kind of have to, in these ecosystem studies, researchers are often trying to figure out what organisms exist. They don't know enough at the start to separate organisms out from each other. So they take samples not from one bacteria or one algae, but from a whole biome of organisms like dirt or water, or the digestive system of a cow. These samples include lots and lots of organisms at once. For sequencing, you pull all the genetic material from all of the organisms out together and work with that. This is a metagenome. To continue our cute metaphor with the bear in the dandelion patch, this is like seeing more bits of the photo. The bear's face, the dandelion patch, the sky above. But first, you sequence all of those pieces separately in those slices I mentioned before. After that, all of those slices still need to be sorted and stitched together correctly. It takes a specific kind of stitching and sorting, but then you can separate out the different organisms that exist in a sample of an ecosystem.

Emiley Eloe-Fadrosh: We have now kind of this whole picture of all of the different populations in the system.

Menaka:Emiley Eloe-Fadrosh is head of the metagenome program at the JGI. So like she said, assembling a metagenome gives you a view of all of the organisms present in an ecosystem. You can see all of those populations no matter how active they are. So this kind of work can pick up all kinds of new or rare organisms. And if you look at many metagenomes,

Emiley Eloe-Fadrosh: You can look at the dynamics of a system in a way that you couldn't do it if you just have kind of these individual pictures.

Menaka: So remember the black bear sitting in the dandelion patch? Imagine stitching together more photos of that same bear over time. In that case, you would see our bear friend not just sitting in a dandelion patch but enjoying a delicious snack of raw dandelion greens and flowers. That's where a time series metagenomic study takes you

Emiley Eloe-Fadrosh: From pictures to movies and from, you know, a single point of a genome to the dynamics of a genome and the dynamics of the community. And so it allows researchers to really kind of ask new questions of the data that they weren't able to before.

Menaka: Like, what conditions cause certain microbes to be more active? Which microbes work together? These are really valuable questions. And you need a few different kinds of expertise to take a look at them. You need sampling and sequencing of course. And then you also need the ability to assemble a bunch of data at once, which actually requires both software and a lot of computing power. The JGI has these available for users.

Emiley Eloe-Fadrosh: I can say this, you know very confidently, there is no other place in the world that can do these types of metagenome assemblies

Menaka: To see why that's possible. I need to take you through a bit of a software history. And instead of starting this story in a month or a year, I wanna start at a crucial place in these computer programs. All of the programs we'll talk about today do this. It's one of the beginning steps of assembling a genome. And I can't say it enough. This first step where we begin is super crucial. It's the kind of step where these programs set up their ingredients. It's like looking through the entire instruction booklet before you start building a piece of IKEA furniture. So here's what happens. Assembly programs start with raw sequence data. They get a bunch of little sequences to sort and stitch together and before they do anything else, they go through all of the sequence data and break it down into littler lengths. Those lengths are called mers, yeah, M-E-R-S. And then the program counts the mers, they do this many times with different size mers each time. So for example, first they break the data into chunks of four and count those four-mers then move on to chunks of five, count the five-mers. Or in genomics terms.

Kjiersten Fagnan: So you do something called k-mer counting.

Menaka: Kjiersten Fagnan is the JGI's CIO. So the K in k-mer counting comes from chopping the data into mers that are K long. K is a number, but it changes. So it's just a placeholder.

Kjiersten Fagnan: So that's where the k-mer comes from.

Menaka: That's very important. Every program we're gonna talk about today has mer in its name because this is foundational, it's where they all start. But it takes a couple of generations of programs to build the capacity to take on the dataset, like the one from Lake Mendota. So let's get to those programs. The earliest parts of this evolution are very punny. We start with a program called Meraculous, spelled mer, of course. Credit to Dan Rokhsar and Jarrod Chapman for designing and writing Meraculous.

Kjiersten Fagnan: And that can run on this like single computer,

Menaka: Which is great. A meracle, if you will. For a while.

Kjiersten Fagnan: You know, if you're just working with like a single human genome or just a handful of microbes, it's not that much data. You transition into plants transition into metagenomes or into fungi. You start to get more and more and more data.

Menaka: Bigger datasets meant more and more k-mers. And then you need memory from more than one computer to manage all that data. And Merraculous wasn't built for that. Suddenly not such a meracle. But there was another program in town that would help.

Kjiersten Fagnan: There was another staff member at JGI whose name is Rob Egan, who had written his own k-mer called Kmernator

Menaka: Like Terminator. Rob wins the award for best software name in this episode. Does he accept?

Rob Egan: Sure. I mean the, that software is dead now. I mean, no, no one uses that anymore, right? And we have much better k-mer counters at this point.

Menaka: Partly that's because around this time, roughly 2010 beyond the JGI, a few more people were looking at this problem.

Kathy Yelick: Hi, I'm Kathy Yelick. I'm a senior scientist at Lawrence Berkeley National Laboratory and also Vice Chancellor for Research and a professor of computer science at UC Berkeley.

Steven Hofmeyr: Hi, I am Steven Hofmeyr. I am an engineer at Lawrence Berkeley National Laboratory.

Menaka: To be clear, Kjiersten, Rob, Kathy, and Steve are all computer scientists, not biologists. They got interested in these genomics problems because they're a chance to program computers to process more useful information faster. A good challenge.

Steven Hofmeyr: When we speed up computers, we add more and more processing units. So to effectively utilize those processing units, you have to do things in parallel. You have to get them all working at the same time.

Menaka: And that might seem straightforward, but it is not.

Steven Hofmeyr: The challenge really is, is it's like trying to make a meal in a kitchen by adding cooks. You know, it's not entirely obvious that you can do it simply by just getting them all to do the same thing. 'cause Then you'll duplicate the work and that's no good. So somehow you've got to communicate and you've got to partition the work and you've gotta take all of these steps to do it efficiently and it's very easy to do it poorly.

Menaka: But they were very qualified to help biologists with the issue of assembling more genomes that included more and more mers. Kathy has spent the better part of her career working to make a bunch of computers do something better together. In particular, she's developed a kind of coding language that manages the communication and divvying up across different computers that makes it possible to compute in parallel on supercomputers. More on supercomputers in the next episode for now. Back to the software. And when did you start thinking about using this kind of work to address biological problems?

Kathy Yelick: You know, that really started with a PhD student Evangelos Georganas. And that was about, gosh, 10 years ago now. And we heard about this problem of genome assembly from a biologist here on campus and also at JGI, Dan Rokhsar.

Menaka: Remember our first program, Meraculous? That's the assembler that Dan shared with Kathy and her team. So Evangelos was taking on the challenge of getting single computer Meraculous to run in parallel.

Steven Hofmeyr: And he mentioned how his work was taking something that used to take a day and getting it to run in a few minutes on a supercomputer and hearing something so dramatic was very exciting. And I thought, oh, I wanna see if I can participate in this project. And that's how I got started.

Menaka: Meanwhile, at the JGI, that programmer that we heard from before, Rob Egan is fresh off writing Kmernator. That program also worked across multiple computers. And he heard a little bit about parallelizing Meraculous.

Rob Egan: Kjiersten was the one who gave a presentation about that just briefly, like, we're working on this. And I asked her if I could, you know, talk to people who are developing that.

Menaka: If this were a superhero movie, we would cut to a shot of Kjiersten and Rob and Steve and Kathy all walking out of a warehouse toward the camera ready to take on their challenge to together. But of course, Evangelos would be there too. And Dan Rokhsar and really a bunch of other people who worked on this together, they'd confidently stride toward their challenge. And that challenge was taking Meraculous, which worked very well at a small scale and making it work at a much bigger scale across many computers.

Rob Egan: Meraculous was a, a bunch of maybe 20 scripts that worked together in a workflow.

Menaka: It was accurate but not efficient. So as a team, they get a new program together. They call it HipMer, that's H-I-P-M-E-R. It's a high performance, so hip, version of Meraculous. Hipmer runs across different computers in parallel.

Rob Egan: But it still had all those same steps that Meraculous had. Because we wanted to be, you know, we produced exactly the same answer that Meraculous could produce. So we could make sure that the this new faster software was accurate.

Menaka: It was. And so far HipMer has stood the test of time. It produces high quality assemblies quickly.

Rob Egan: So we still run it. And the plant groups, especially HudsonAlpha, still run it and they need they want, they want to run, I think 50 agave plants over the next couple weeks.

Menaka: But HipMer is not the end of this story. Because remember all of this software optimization is happening in California. Meanwhile, in Wisconsin there are graduate students out on a lake. They are sampling the microbial communities of Lake Mendota. And their dataset is not a single genome dataset. It is a metagenome dataset. And there are more and more metagenome projects like it coming online.

Kathy Yelick: So that was one of the reasons to try to say, well, how do we get more computing power into the problems that come from the data that's coming outta the sequencers? 'cause You could tell there was this data tsunami happening with all this data coming from the sequencers.

Menaka: After the break, our team tackles that data tsunami of metagenomes.

Allison Joy: The JGI supported this project via the Community Science Program. This program provides genomic resources for projects with Department of Energy relevance. And we accept proposals from scientists at all career stages.

Menaka: Usually writing a research proposal means requesting support for a project in the form of money. But the JGI is not a funding agency. We're a user facility, so an actual lab in Berkeley, California with all kinds of sequencing and 'omics and bioinformatics capabilities. So proposals at the JGI work a little differently.

Dan Udwary: You know, we don't give out money. Instead we give out capacity and do the work that you need done, right?

Menaka: And users don't pay for that work. It's funded by the Department of Energy.

Allison Joy: You can find out more about submitting proposals to the JGI on our website, head to joint gino.me/proposals. We've also got links waiting for you wherever you're listening to this episode. Either in the episode description or the show notes.

Menaka: This is Genome Insider from the JGI. To recap where we've been so far, we've made our way through three iterations of programs. First there was the assembler Meraculous and a k-mer counter, Kmernator. Then a team of brave software engineers rewrote Meraculous so that it would run across multiple computers. And we arrived at HipMer. That's a lot of progress, but we're headed toward a 25 terabyte dataset from Lake Mendota. There are still a few more steps before processing that metadata is a reality for this team working on it. Here's Kathy Yelick again from Berkeley Lab.

Kathy Yelick: We always had the interest in doing metagenome assembly, because we saw this as a place where supercomputing was especially important because the data sets in environmental genomes become so large.

Menaka: And that's because metagenomes sample all of the organisms in an ecosystem at once. Steve Hofmeyr, also from Berkeley Lab.

Steven Hofmeyr: But in a metagenome you've got a mix of genomes, potentially thousands of genomes in your sample. And they occur at different abundance, different frequencies. And ideally what you want your assembler to do is to pick out each of those genomes separately and assemble them each individual.

Menaka: It's a very complicated problem.

Kathy Yelick: I sometimes say the metagenome assembly problem is like somebody has come to your house with a dump truck and dumped puzzles, thousands of puzzle pieces all mixed together from thousands of different puzzles all onto the ground. And and now you're trying to put them together.

Menaka: So remember ,getting to HipMer, this team essentially created an algorithm that would do the same steps as the previous one, just in a more efficient way. Here's Rob Egan from the JGI.

Rob Egan: So then we tried to put that, that framework into a metagenome context and things just kept falling over because it was just a little bit too big. Even though we were distributed, the software wasn't set up. It was very steppy,

Menaka: Following the exact same steps, meant lots of reading and writing for a computer. So it was slow and crashy. They opted to set up a new program to run a different way. It runs in memory rather than reading and writing. And it loosens the algorithm steps to let in a little more randomness into the assembly.

Rob Egan: But that makes us go a lot faster because we don't have to make all these assumptions. We don't have to sort these datasets before we do it. We just know that we're gonna get approximately the same thing every time.

Menaka: That program became MetaHipMer. So if you're keeping track, that's high performance Meraculous for metagenomes.

Kathy Yelick: Then once we had a metagenome assembler, it has repeatedly assembled the largest metagenome dataset ever. So it just keeps kind of beating its own record, if you will,

Menaka: Until the datasets got even bigger.

Rob Egan: And we ran up to the limits of what we could really do with that software.

Menaka: Time for another revamp. Our last one, at least in this episode.

Rob Egan: In like 2019, 2020, we rewrote HipMer or MetaHipMer into MetaHipMer2. And that was complete rewrite of build code. So we took the algorithms and mostly ported them, but we rewrote the entire set code base so that it's now much faster, much more scalable and much more efficient.

Menaka: MetaHipMer2 is where this software meets the Lake Mendota dataset, which was actually still big enough to test this program.

Rob Egan: That was a really great challenging dataset. So we, we would kind of titrate in the data to see where we, where we break.

Menaka: They'd feed in bits of the data to check that it was working stepwise before running. First a 10th, then an eighth, a sixth, a fifth, and then half the data

Rob Egan: Until we, we were satisfied that that could work. And then we do the whole data set and break again. Then we try it again, rejigger the computer or rejigger the algorithm a little bit and then make it work.

Menaka: This has always been sort of a try, try again, keep trying kind of project.

Rob Egan: Throughout this whole process, we've been kind of making it more stable and more robust to larger and larger datasets

Menaka: That's important for this Lake Mendota dataset, which we're headed back to. But it's also important for the scientific community in general.

Kathy Yelick: Once you have a tool that can analyze tens to hundreds of terabytes of data, then people have new scientific questions they wanna ask and they go out and collect bigger data sets.

Menaka: And some of that work benefits, very small stuff.

Steven Hofmeyr: And if you think about the very rare microbial genomes that occur in those samples from one sample, there's probably so little that you can't really reconstruct that rare genome.

Menaka: In other words, if you only assembled one piece at a time, you'd completely miss those very rare organisms. They'd only show up in tiny bits where honestly they'd just look like errors.

Steven Hofmeyr: But once you take multiple different samples and you compose them together with this co assembly,

Menaka: That's what MetaHipMer and MetaHipMer2 do. They co assemble lots of samples all at once.

Steven Hofmeyr: You suddenly have enough quantity at that low level that it no longer looks like an error, as you were saying, but now it looks like a proper genome. And we've sort of demonstrated that it's possible to extract very rare genomes using this approach that you could never get if you didn't do these very large scale coassemblies.

Menaka: And you can coassemble sequences that have already been analyzed, actually. So this could add entirely new layers to our understanding of ecosystems.

Steven Hofmeyr: But it also is very valuable because it forces systems to sort of broaden their perspective about what hardware should do and could do. You know, because these are very important problems. Large scale metagenome assembly is important. It's not something that should be just disregarded because we don't have the technology to do it.

Menaka: This is a case where many things advanced at once. Sequencing technology, the software development, and then of course supercomputing hardware. And bridging a collaboration across these fields has meant some really exciting metagenomic results. Here's Emiley Eloe-Fadrosh, head of the JGI's metagenome program, again.

Emiley Eloe-Fadrosh: I view it as a really successful collaboration because you have some of the, you know, world experts in computing sciences who are interested in hard biological problems

Menaka: To develop MetaHipMer2, this collaboration became part of the DOE'S exascale computing project, basically using supercomputers to tackle biological problems. They called this the ExaBiome Project, since it's focused on microbiomes.

Emiley Eloe-Fadrosh: I hate to say visionary, but I think it was visionary . Just because I think visionary is kind of like overused, but I think it was very visionary of the ExaBiome team to think ahead about how can you support metagenome assembly for doing really, really large scale projects like this. And then Mendota came online and it was like, oh, this is the perfect data set.

Menaka: The Lake Mendota dataset that we're working our way back to, that's ultimately 500 metagenomes in 2021. It was the biggest dataset that MetaHipMer2 had ever assembled. And it's much bigger than what Emiley imagined working on when she started out as a scientist.

Emiley Eloe-Fadrosh: Yeah, no, definitely not. I remember in grad school I was working on an isolate in the lab, but I remember we were sitting out during lunch talking with this visiting scientist and my advisor was like, oh yeah, well Emiley is working on one of the first metagenomes from this, you know, really unique deep sea environment. And this is just like the beginning of metagenomes. And I anticipate that all of my future graduate students will be doing, you know, much larger metagenomes and what, and I just remember sitting there feeling like, yeah, that's, that's interesting. I don't know what to think about that .

Menaka: But now of course, a few decades on that prediction has come true and it's exciting for Emiley and the whole field.

Emiley Eloe-Fadrosh: I think for the field itself, the pace at which all of these, you know, kind of computational developments and then how sequencing has just massively gotten cheaper and higher throughput. It just has opened up how research can be done.

Menaka: And at the JGI, this work is the result of a 25 year history where sequencing has galloped along and people have assembled bigger genomes and bigger metagenomes every year. The Lake Mendota dataset is just one of those behemoths, but before we get to the lake, we'll see how these data make their way through an assembler like MetaHipMer. And that is a super computing story that's in two weeks.

Menaka: So again, that was Trina McMahon from the University of Wisconsin at Madison, Emiley Eloe-Fadrosh, Kjiersten Fagnan and Rob Egan from the JGI and Kathy Yelick and Steve Hofmeyr from Berkeley Lab. You can find out more in our episode description. This episode was written, produced, and hosted by me, Menaka Wilhelm. I had production help from Graham Rutherford, Allison Joy and Massie Ballon. We had music in the middle of this episode by Cliff Bueno de Mesquita, who's a multi-talented postdoc at the JGI. If you liked this episode, help someone else find it, tell them about it, email them a link, or leave us a review wherever you're listening to the show and subscribe so you don't miss future episodes. Genome Insider is a production of the Joint Genome Institute, a user facility of the US Department of Energy, office of Science, located at Lawrence Berkeley National Lab in Berkeley, California. Thanks for tuning in. Until next time.