Genome Insider S3 Episode 4: From Sample Shipments to Sequences – A Tour of the JGI's Sequencing Pipeline

Menaka: Today, we're taking a look at a part of the JGI that's central to all kinds of projects – but also, a place very few people really ever get to see: our genomic sequencing lab.

As a user facility, the JGI supports lots of researchers' work every year. And for most of those users, working with the JGI starts out as writing a project proposal.

Then, there are samples to collect – that part takes users all over the world! To forests, and lakes, farms, and of course, occasionally, Antarctica.

These samples generate troves of genomic data, to drive all kinds of science. And for most JGI users, the first step in that data creation pipeline sounds like this – sending a box to Berkeley.

They hand off samples in a shipment to the JGI, then, after some time passes, get their data back, online. But there's a lot that happens as samples go from shipment to sequences.

So - today, we're headed into the lab, for a look at how that goes.

Menaka: This is Genome Insider! Where we dive into the JGI-supported genomics research that could advance clean energy and protect our environment.

I'm your host, Menaka Wilhelm.

And like many JGI users, I'm new to the JGI's genome sequencing lab. Chris Daum is showing me around.

Chris Daum: So I'm Chris Daum. So I lead the sequencing platforms group here at the JGI. My team is specifically tasked with the automated sample prep of sequencing libraries, and then loading those samples onto one of our sequencing platforms, the Illumina, or PacBio sequencers.

Menaka: And before we get to sample prep and sequencers, let's talk a little bit about sequencing at the JGI. Reading DNA's nucleotides is how the JGI got started. That was 25 years ago, when a complete human genome didn't exist yet. The JGI was formed to help sequence the DNA of three chromosomes for the Human Genome Project.

Since, then, sequencing has changed a bunch, and the JGI has expanded to do a lot more, too.

Chris Daum: We're not just sequencing instruments.

Menaka: For many projects at the JGI, a sequence is really just the beginning. JGI users also get support understanding and annotating genomes, and looking at which genes are turned on or off, or which metabolites are being produced, under certain conditions. For some projects, the JGI will even write, or synthesize, DNA to test specific questions that a genome brings up.

Chris Daum: So it doesn't end once we generate the sequence. We take it to completion with the analysis for projects as well.

Menaka: But many projects do start with a sample to be sequenced. So that's where we're starting, too.

Sequencing samples is almost a bit of a ballet – many moving parts and precisely set steps. You can think of it happening in three acts.

Act 1: Sample Management. So, basically, receiving samples and storing them carefully, until sequencing starts.

Act 2: Quality check and library creation. So that's checking to be sure that samples meet requirements, then preparing them for sequencing.

Act 3: Sequencing. So running samples through sequencing machines that read their basepairs.

And at the JGI, this sequencing is a high throughput operation. Every year, our labs process somewhere around 35,000 samples — so all of these steps happen on a giant scale, with hundreds of samples running in parallel, thanks to automation. At any given time, hundreds to thousands of samples are in progress. So let's see what that looks like.

Chris Daum: Happy to yeah, show you around.
Act 1: Sample Management
Menaka: The JGI handles samples from far and wide — we could be talking algae from the Arctic, or gut microbes from Australian camels, or even giant bacteria from the caribbean.

Chris Daum: So a lot of these times we're looking at these, these organisms that are kind of extreme in their nature. So we're trying to figure out like, how do they thrive in different environments? And so they come from very interesting places for sure.

Menaka: But when those samples make their way to the JGI, in Berkeley, California, they all arrive the same way.

Chris Daum: We actually ship pre-labeled containers, to the PI and the user that they'll put their samples into.

Menaka: Those labels have barcodes. So when samples arrive, they're all ready to be welcomed in.

Chris Daum: So this is our sample receipt lab here. Typically samples are sent on dry ice or wet ice. And here, they get unpacked.

Menaka: Most samples don't get sequenced right away. So frozen samples are headed straight back to cold temps for storage. There's a very clear star of the show in this sample receipt lab. The freezers! There are two. Each one is a giant, walk-in closet-sized machine. And these freezers have names. They're SAMs.

Chris Daum: So it stands for sample access manager and that's just the vendor that makes them, so that's their name for 'em.

Menaka: These SAM freezers are very different from your standard lab freezer. They don't have handles or doors! Instead, they have computer monitors, and square tinted windows at eye level.

When you go to put samples into this freezer, you load tubes or racks into a little black compartment. Then, the SAM scans their barcodes, and moves the samples into their proper space for storage.

The idea with these freezers is to keep users' samples stable, secure, and organized. Many of these specimens are totally irreplaceable. So these SAMS keep everything at a super consistent minus 80 degrees Celsius — no temperature shifts from open doors. And with a machine moving everything around, they're super organized — something not every lab freezer can say.

Chris Daum: So we have a computer tracking system where we check all the IDs of all of our samples, as they flow through the lab.

Menaka: OK and that noise that just interrupted Chris a bit – that's also from the way these SAMs move samples around.

Chris Daum: The sound that you just heard there is part of the mechanical, robotic system. And so the robotic arm uses air pressure to move it around, and that's just the off gassing.

Menaka: It's all totally automated. Including when it's time to pull a sample for sequencing.

Chris Daum: So you can think of these SAM freezers as automated vending machines for samples.

Menaka: You order up your sample from a computer system and the SAM actually finds it for you. We ran into someone doing just that –

Kathleen Lail: So what's happening is I'm just pulling out a single tube right now — Hi, I'm

Kathleen Lail and I work in the sample management group.

Menaka: So Kathleen has ordered up a single test-tube. It'll take a minute for the SAM to retrieve it, because it's got to locate the sample, then move it out. Which also involves shifting around other tubes.

Kathleen Lail: So that it, keeps the most open spaces. And it has to create an open rack to put this tube…

Menaka: And then, voila – sample delivered! Which is where we start that second act - checking and preparing a sample to be sequenced.
Act 2: Quality Check and Library Creation
Menaka: Setting up for sequencing involves plenty of quality control. First, technicians measure how much DNA made it into a sample. The sequencers need 200 nanograms of genomic DNA to make JGI's standard PCR-free whole genome shotgun sequencing library. 200 nanograms might sound very small, but it turns out to be a lot of genetic material – hundreds of trillions of base pairs. And sequencers can't read that DNA in its original form.

Chris Daum: So if you think of DNA when it's wound up in a chromosome or it's actually within a cell, we're talking millions of nucleotides in length. And so when you do an extraction, if you do a really good sample extraction, you can actually recover very long strands of DNA from a cell. But that's too long for us to work on.

Menaka: So no matter what kind of sequencing you're doing, there's chopping involved.

Chris Daum: So you take these long strands, long molecules that are hundreds of thousands or even millions of nucleotides long, and then we chop 'em up into smaller pieces and then we're able to sequence those individual pieces, uh, much more easily.

Menaka: This step chopping DNA down to size, so a sequencer can read it is called library prep.

Chris Daum: So it's really just that analogy of, you know, a library, a book contains information, so does, uh, genomic DNA that if we sequence it and read it just like a book, it gives us information about the sample.

Menaka: And for many samples, library preps happen with the help of more automation. So there are robots up ahead. First, a quick, quick break.

Menaka: If you're a researcher who'd like to send samples into the JGI, for a project focused on clean energy or the environment, one way to do that is to submit a proposal to the JGI Community Science Program.

We have projects in genomic sequencing, DNA synthesis, and metabolomics in all sorts of non-human organisms.

Our current community science program proposal call is looking for projects in Functional Genomics. That call closes at the end of January — and we'll have an episode soon with more details on submitting a proposal — but in the meantime, you can find more information at our website. There's a link in the show notes.

Menaka: And now, back to our tour of the JGI's sequencing pipeline. So far, we've seen how DNA samples come in. They're stored in big, automated freezers until it's time for sequencing.

Then, the next step is setting up samples for sequencing – so quality checks, and chopping samples into bits the sequencers can read. And because the JGI works on a really wide range of projects — on plants, and algae, fungi, viruses, bacteria, achaea, you name it — this lab has really tailored capabilities for studying all of those organisms.

Chris Daum: We have, you know, several dozen, I think close to 40, different types of sample prep that we do, depending on the sample type, the type of sequencing that we're doing, the type of data analysis that we're trying to do as well.

Menaka: And in terms of setting up samples, there's another important piece of this pipeline that affects sample preparation. The JGI does two different kinds of sequencing, and samples get prepped differently, depending on which one you're using.

There's long read sequencing, where a machine — from the company PacBio — reads hundreds of thousands of base pairs at a time. And there's short read sequencing, where an Illumina machine reads bits of DNA that are hundreds of base pairs long.

Chris Daum: And then the two technologies, they're used for different applications.

Menaka: We'll hear more about the long and short of that soon — for now, we're headed back to our library prep step. And because different sequencers read different different lengths of DNA, library prep happens differently for each machine.

Chris Daum: So for a PacBio library,

Menaka: That's the long-read sequencing,

Chris Daum: We don't fragment the DNA quite as small, so we keep it longer on purpose just so we can sequence those longer reads.

Menaka: So technicians do these long-read library preps by hand.

But short-read sequencing happens on a bigger scale. So those library preps happen in a 384-well plate, with help from machines.

Chris Daum: So we have these liquid handler robots,

Menaka: And these robots sit right on top of a lab bench. They look like big, black boxes, with cool technological silver accents. Inside, they've got a motorized arm that moves side to side, with basically, a whole grid of pipette tips on its end.

So a robot like this is all set up to do benchwork on its own — from pipetting, to scheduling wait times, and of course, changing pipette tips in between. And DNA library preps require all of these capabilities.

Chris Daum: So for a DNA library prep, we're going from extracted genomic DNA to a sequencing library. And to go from that starting material to that final material, there's a series of enzymatic steps, or chemical reactions that happen, to convert that genomic DNA into a library that can be sequenced.

Menaka: And this liquid handler carries that process out in each well of this plate. It'll pipette in the enzymes to chop DNA down to size, and add the template DNA that'll allow a sequencer to read these samples.

Chris Daum: And what this allows us to do is a single technician can, instead of just processing a few samples at a time at the bench, they can use these automated robots to help them process hundreds of samples at a time.

Menaka: So that's a big way this lab accomplishes such high throughput – and it wraps up this library prep step.

Chris Daum: Next would be the sequencing step, and we do perform one more quality control step. So once the libraries are created, we again need to quantify them to know the concentration of our libraries before we can load them on the sequencer.

Menaka: So – once samples are checked twice and prepped in their libraries, it's time for the grand finale.
Act Three: Sequencing
Menaka: Which brings us to our sequencers!

Chris Daum: So we have the one NovaSeq6000, and then the three PacBio Sequel2E Instruments,

Menaka: They are named Instrument one, and Instruments 1, 2 and 3 — but here I would like to take a quick tangent into the past, because in the mid-nineties and early 2000's, the JGI had far

more instruments than this, and so they used to have names.

The sequencing labs in Walnut Creek, Bldg. 100, around 1999 (Photo courtesy of Lawrence Berkeley National Laboratory)

Chris Daum: When we had lower throughput sequencers, we required a lot of instruments, we would name them. So this was back in like the earliest Sanger sequencing days, capillary sequencing, so at our old facility in Walnut Creek, we had four different labs. There were just rows and rows of sequencers, and I think we had just over a hundred instruments at one point. And so I do recall that, you know, we had a series of sequencers that were superhero names, so like Marvel and DC superheroes. And then we had another set of sequencers that we used, a University of California and Cal State University mascots to name them. So that's certainly been part of the past, but that usually comes with it when you have a lot of instruments and you're trying to keep track of 'em all.

Menaka: Since then, sequencing technology has shifted enough that while the JGI used to work with four labs of sequencers — we now operate with four very powerful sequencers, at the other end of this one lab.

Chris Daum: And they are a little bit louder with the fans. Yeah.

Menaka: So first, let’s talk short read sequencing. This all happens on one machine.

Chris Daum: So this is our NovaSeq instrument. So this is our flagship Illumina sequencer. It's the primary data generator that we have here at the JGI.

Menaka: It’s got a fairly similar form factor to an extremely streamlined xerox copy machine, with a monitor ready to show the level of sequencing happening. Samples go into the NovaSeq machine in a flow cell — so picture a glass rectangle, roughly the size of a smartphone screen, bordered by a thin, white plastic rectangle.

Chris Daum: So this is the Illumina flow cell, and this is really the heart of the sequencing system.

Menaka: The glass rectangle at the center of the flow cell is two glass slides, so samples flow in between them. And they’re split up into four chambers that run the length of the glass slides — it looks a little like four lanes in a mini swimming pool. Those chambers are where sequencing libraries flow in.

Chris Daum: And within the chambers there's surface oligos, which are short strands of DNA and they're complimentary and sequence to those adapters that we've attached to 'em.

Menaka: All DNA bits get template sequences as part of the library prep,

Chris Daum: And so the templates, um, as they flow through they'll hybridize onto those oligos and get captured to the surface of the flow cell. And once they're captured, then the sequencing process can start.

Menaka: Are they running right now?

Chris Daum: It actually is running right now, so you can see here, two runs are going, they're about two thirds complete based on the status bar. So this one is projected to generate just under four Terabases of sequence data.

Menaka: So that's 4 trillion base pairs – And sequencing this kind of sample works differently than sequencing human genomes. But for an idea of size, our human genomes are about 3 billion base pairs — so this sample run is working with the number of base pairs contained in over a thousand people's genomes. And that's on the small side for this machine — it can generate 8 terabases of sequence data in one run, which takes about 2 days. And this kind of sequencing fuels lots of different projects.

Chris Daum: So if you're just doing a re-sequencing study or you're just mapping to a reference genome, short reads are perfect for that — if you're just looking for SNPs and things or mutations.

Menaka: So that's short read sequencing, on the Illumina NovaSeq machine. Then, there's the three PacBio machines — the Sequel IIe's. They're roughly the size of a refrigerator, with a black top half and a grey lower half – and they do long read sequencing.

Chris Daum: So you get these really nice long sequencing reads that are then very easy to assemble, to get really high quality complete genomes out of them. And they can also do full length RNA seq, transcript sequencing as well.

Menaka: And to go into a PacBio machine, samples travel in a different vessel.

Chris Daum: Yeah. So PacBio, they use what they call a smrt cell is what they're using. So that's s m r t, and that stands for single molecule real time, sequencing.

Menaka: And this long read sequencing is lower throughput — you get 5 or 6 million sequencing reads, but the reads are much longer — hundreds of thousands of base pairs.

Chris Daum: So they're really kind of these, there are these nice complimentary technologies. So one is very long reads, just not very many of them. The other one is lots of reads, but they're short.

Menaka: And whether it's been through long-read sequencing or short-read sequencing, a sample that's come this far has been through quite a pipeline. From the SAM freezers, to library prep, and finally a sequencer. And in all of these steps, skilled scientists are making sure

everything that comes through the pipeline will be high quality and accurate. And those sequences are really not the final step!

Chris Daum: So we have really expansive bioinformatics teams and groups that will work with our users, to help do the analysis that they're after.

Menaka: So while our tour ends here with samples running through sequencers, for most projects, this part is really just the beginning.

Menaka: That's it for our spin through the JGI's sequencing lab. We'll be back in a month — with more details on submitting proposals to work with the JGI!

This episode was written and produced by me – Menaka Wilhelm, with production help from Massie Ballon, Allison Joy, and Ashleigh Papp.

Many thanks to Chris Daum, Kathleen Lail, Tanja Woyke, and Len Pennachio!

If you liked this episode, help someone else find it! Tell them about it, send a link over, or leave us a review wherever you're listening to the show.

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!