



# Q&A: Dana-Farber's Paul Morrison on Running a Helicos Sequencer in a Core Facility

October 13, 2009

**Newsletter:** [In Sequence](#)  
[In Sequence - October 13, 2009](#)



**Name:** Paul Morrison

**Age:** 54

**Position:** Principal research scientist (since 2004) and director, Molecular Biology Core Facilities (since 1992), Dana-Farber Cancer Institute  
Associate director, Molecular Virology/Biology Core Facility, Center for AIDS Research, Harvard University (since 2004)  
Principal associate, biological chemistry and molecular pharmacology, Harvard

Medical School, since 1996

**Experience and Education:**

Laboratory supervisor, Dana-Farber, MBCF, 1987-1992

Research technician, Harvard Medical School and Dana-Farber, 1980-1987

Electron microscopist technician, Marine Biological Laboratories, Woods Hole, 1979

BS in botany, University of Massachusetts, Amherst, 1978

**Paul Morrison** is the director of the Molecular Biology Core Facilities at the Dana-Farber Cancer Institute, where he and his colleagues provide a variety of services, including DNA sequencing, protein mass spectrometry, and DNA and peptide synthesis.

Earlier this year, Morrison decided to take a pass on the first crop of high-throughput sequencing platforms. Instead, he brought in the Helicos Genetic Analysis system — the first and so far only commercially available single-molecule sequencer — which he thought might be a better fit for a core facility.

*In Sequence* spoke with Morrison recently to find out about his experience with the system, and how it compares to other platforms. Below is an edited transcript of the conversation.

**What kinds of services do you provide at the Molecular Biology Core Facilities at Dana-Farber?**

A whole suite of services. Back in 1985, I bought one of the first DNA synthesizers from Bill Efcavitch [now Helicos CTO] [and colleagues] at ABI, and today we provide peptide synthesis, protein sequencing, and a whole lot of mass spectrometry. DNA sequencing has been a large part of the services starting from 1990, when we bought one of the first fluorescent DNA sequencers from ABI. During the last few years, we have been concentrating on mass spectrometry. With next-generation sequencing, I have been staying on the sidelines.

### **How did you decide to bring in the Helicos platform, as opposed to more established platforms, like the Illumina GA?**

Before even the 454/Roche platform was commercially available, I was talking to [the company] about that instrument, and when Solexa brought out their instrument, turning into the Illumina GAII, I was certainly chatting with them, and with ABI about the SOLiD.

I took a pass on every single one of them. One factor was, the Broad [Institute] is very nearby. If they purchased a lot of Solexa/Illuminas, then I didn't have to. They have the front-end expertise to make those machines run efficiently with a lot of people and a big pipeline, [and at] the back end, they could very efficiently handle the informatics. Looking at it from a core facility viewpoint, I would rather have consulted with researchers at my institute and told them, 'Go over to the Broad to do that project; they can do it more efficiently.'

I have been chatting with Helicos for two and a half to three years now, and right at the beginning, I could see that the Helicos much more suited my idea of how to do next-gen sequencing in a core facility. The fact that it did 50 samples at a time was very attractive — that would hopefully get rid of a backlog [of samples] without having to buy a lot of instruments.

But the real killer — and it's become more obvious now that we have been using it since April — is that the front-end preparation is really far, far simpler than the other instruments, because there is no amplification, there is no library building.

Also, I could get the data to labs quickly, because the file size for the Helicos is only 2 gigabytes per sample, and people can download the data from the FTP server that we have in minutes. And, in fact, we have learned pretty easily how to filter the data ... so the labs that we are handing that data to can immediately start getting some scientific [results] out of them.

### **How long have you had the Helicos instrument?**

It rolled out of the elevator Friday, the 13th of March, and we began our first run of our own real samples — after Helicos ran a couple of sets of their samples to test it out — on the 13th of April.

### **Are there any requirements for the instrument that are different from other lab equipment?**

We added a \$38 house fan that we put up over the refrigerator so it would blow air over to the thermostat. I had a spare room that had mass spec equipment in it up until a month or two earlier, so it was a well air-conditioned room. If the instrument was going to be photographing very, very small objects day in and day out, and they needed the coordinates to be exactly the same, I assumed that if I could have a room that stayed at a very constant temperature, I would be better off. And with that fan added to that room, the temperature stays spot-on. It's a pretty

solid building that we are in, so we didn't have to do anything else.

### **What have you used the instrument for so far?**

Even before we signed up for the Helicos collaboration, I wanted to make sure that we had enough samples to run it all summer. So Zach Herbert in my lab and I sat down, and we came up with a list of 15 labs in the institute that we thought had a significant amount of next-gen sequencing experience — most of them had received 454 or Solexa data already — that they might be interested in Helicos data. So I sent those 15 labs an e-mail saying, 'We are going to do this project through the summer with the Helicos, would you like to be in on it? If you do, you have to send me \$13,000 immediately. If you do that, you then get to run samples all summer long.'

Of the 15 labs that I asked, 12 signed up within 24 hours. That was a good thing; I got a lot of money upfront, so I could feel comfortable spending [it on] chemistry and reagents throughout the summer without feeling like I was digging myself a hole.

For those 12 labs, we have generated data from 500 samples from April 13 to today [and are currently running another 50 samples]. Of those samples, probably 80 percent were ChIP-seq runs, and of the 20 percent left, most of them were digital gene expression, or resequencing of small genomes like *E. coli*, or *C. elegans*.

### **Why such a high percentage of ChIP-seq?**

I work at a cancer institute, and all of these people are interested in gene regulation. More than one gene — they want to look at the whole picture, and there is nothing that gives you a better genome-wide profile than ChIP-seq, where for their protein or pathway of interest, we are now giving them more data than they could ever have thought possible, and they are happy about it.

### **How have you validated the data? Have you compared it to that from other next-gen sequencing platforms?**

These 12 labs are not going to jump on a bandwagon unless they have proof that it's better than what they have been using. So probably the first 150 samples were samples that had already run on Solexa or 454, and I would especially like to thank Myles Brown [of Dana-Farber], and Tom Westerling in his lab, for extensive comparisons to the other platforms.

[With] the single-molecule sequencing of the Helicos, the removal of the amplification during the sample prep is not just for convenience's sake during the sample prep, but it's huge for removing all the amplification bias that can occur when you create libraries on any of the other platforms that require amplification.

There are a lot of ways to get amplification artifacts, not just by GC bias, or by jackpotting fragments so that the signal goes up; there is a lot going on in PCR. Especially if you are using a very small amount of DNA to start — and with ChIP-seq, we normally use 3 to 4 nanograms of unamplified DNA that goes into the Helicos — you can be creating a lot of noise, a lot of false positives that require input controls, which require repeating the experiment, or, if you do the experiment a month later, duplicating runs. Whereas with the Helicos, the data is a lot more

solid, a lot cleaner.

Using the word "a lot" is not really quantitative, but with 500 samples, I'm feeling very comfortable that we are handing our researchers better-looking data than they have seen before.

### **How does the Helicos compare in terms of time-to-result?**

The fact that the sample prep is shorter [than for other platforms], and the fact that I can run 48 samples in 10 days, all in all, I can get the data back to [researchers] a lot faster. In other places where they are getting Solexa data, [our researchers] are waiting a little bit longer.

They are also being charged around \$1,200 for a Solexa run, whereas we, back-of-the-envelope, have decided that we are probably going to never go above \$500 per sample for a Helicos run.

### **What costs does the \$500 per sample include? For example, does it count in instrument amortization?**

It doesn't include the amortization of the instrument, because I'm hoping that a grant is going to pay for that. If we buy a normal quantity of chemistry from Helicos, it cost \$325 per sample for the chemistry and incidentals, like streptavidin chips and a few other things. That leaves \$175 per sample to be used for service contracts, full-time employee [costs], informatics, server expansion. It's some \$230,000 that I would need to cover FTE and server [costs], and at \$500 [per sample], I've got it covered. At \$500, I could even cover the amortization of the instrument.

### **So it sounds like it's a viable option for a core facility?**

We had 48 billable [samples] each time we ran the instrument, and there were very few failures where we would say, 'That was our fault, we'll run it again.' I've been doing core facility services since 1986, and financially, [the Helicos is] by far the easiest thing to figure out, that it would be very easy to cover [the cost] at \$500.

### **Do you get the same amount of data per sample, compared to other platforms?**

Actually, I think we get quite a bit more. With a Solexa run, you need to run a software program that removes all fragments that are exactly identical, and then you come up with a number of unique fragments in a run, and I think generally, 4 to 6 million is about what people are getting. We don't have a Solexa, so I'm just taking my researchers' word for that.

With the Helicos, there is no such thing as running a program to remove exact duplicates because there aren't any because there is no amplification. We are getting, on average, around 14 million unique reads. They are shorter than the Solexa reads, but in ChIP-seq, the number of unique reads is the big deal. As long as they are long enough so that they find where they originated from, they are counted as a tag — they may be 30 base pairs in length and they are a perfect tag.

### **Would the read length matter more for genome sequencing applications?**

For the Helicos as a tool, I don't think full human genome sequencing is the right thing for it to do.

I think ChIP-seq is, today, in 2009, the perfect thing for it to do. Then after that, RNA-seq, where you are generating one cDNA copy of the messenger RNA.

Actually, the second-most popular application here has been digital gene expression. We liked that one because it's as simple to do as ChIP-seq for us — you poly-A-tail, you quantitate, and you load it on the Helicos. On the informatics side, for us, it's even simpler, we don't do any mapping, we have a small pipeline that creates a very large spreadsheet, and it just tells you how many counts for each gene in the genome have occurred.

**In every talk Helicos gives at conferences, the platform's raw read error rate comes up. Is that a concern at all?**

It has a higher error rate than Solexa. But a vast majority of the errors are deletions, where the Helicos did not perceive a flash of light. With ChIP-seq, and with digital gene expression, if you have, on average, a fragment size of 35, and you have one somewhere, in every single one of those fragments, that's perfectly OK, they are still unique tags.

Early on, I selected labs that had a lot of informatics experience, so the first thing they did was, they took the Helicos data and they put it into their Solexa informatics pipeline. And what they got out at the end was very, very few valid reads. And that was because the Solexa pipeline is tuned for Solexa reads that are uniform in length, much longer, and have a lower frequency of errors.

As soon as I got them to use the Helicos software for the filtering, which understood that the sequences are of variable length, and that there will be a 4 to 5 percent frequency of error, but most of them deletion errors, the number of unique reads went way up. And when you used those valid reads and mapped them to the genome, most of them stuck. So the Helicos pipeline was passing many more valid fragments. I have a feeling that some people may have fallen into that same trap. They had Helicos data, and they ran it through their standard pipeline for another next-gen sequencer, and they got the wrong results because they were not using the Helicos software. Even though all the Helicos software is open access, anyone can download it, it does take a chunk of time, and certainly, you don't want to do it if you don't have to. But if you think that you can use the software that's available, you are making a big mistake.

**In terms of software, is it a user-friendly platform?**

In the beginning, no, but very quickly, yes, we have found the software to be user-friendly. Mostly because Helicos is very open to [helping us.] It didn't take very long to get a ChIP-seq pipeline and a digital gene expression pipeline up and running.

Zach Herbert in my lab, who supervises all the genomics projects of the facility, wanted to know everything about how the Helicos works from top to bottom. He's been the only one who has done the sample prep and the informatics. The software has been relatively painless to use. Sometime in the future, when the Helicos is an extremely popular platform, and the whole informatics community goes crazy about it, I know that the pipeline will be a lot smoother than it is now. But the way it is today is completely usable, though I think Zach would not say it's easy to use. When we got the instrument in March, I showed him his first Unix command. He knows a lot more Unix today, and he is running all of the Unix code to filter and map the data. It's something

that the had to learn over a period of a month or so, but it's nothing too scary.

### **How robust have you found the instrument to be?**

I have done 550 samples, 11 runs, and there was only one run that had an error. When you do a Helicos run, there are two 25-lane chips running at the same time, and the software just decided that one of the chips was done [early on], so it stopped and that one chip failed. Helicos knew that night that the failure had occurred because they are monitoring the instrument from their place. They told us in the morning that one side has failed but we should continue because the other side was still running. They came in the next day, ran a bunch of software and updated a couple of modules, and it hasn't happened since.

Inside this very large 1,500-pound box, the Helicos is doing a lot of very complicated chemistry, a lot of deliveries to a very small space. And then you look at the optics for it, it can take a photograph on a Monday of a very small spot, and then they come back seven days later and photograph the same area, and it has to be in the same register within that very small spot or else all the data is trash. It's taking a lot of photographs, and then it processes all that data, and what we get out is just the 2 gigabyte files of the 20 to 40 million reads that we then filter. It's doing a huge amount of work, and I'm somewhat surprised that it hasn't broken down more than just that one time, and that was a just a software glitch, it wasn't a mechanical failure. For being a dependable platform, knock on wood, the one that I got has been working like a champ.

### **I understand that Dana-Farber has had the instrument for a trial period. So based on the results, you have decided to purchase it?**

Yes, but what I want to do and what I'm able to do are sometimes two different things. What I really want to do is have two Helicos, and I would purchase them. I have a high-end instrumentation grant in, and if that is funded, I am immediately writing out a Dana-Farber check for one Helicos, and I will immediately try to figure out a way to pay for a second one.

### **How do you expect the platform to improve in the future?**

The RNA-seq protocol has been getting more robust through the summer, and I expect that to continue. The paper that was just put up by Helicos of the straight RNA sequencing (see [\*In Sequence\* 9/29/2009](#)), is something that would certainly be very attractive to do. I could see how quite a few researchers in my institute would like to jump on that.

The future change of the Helicos that I'm most excited about, though, is the one that Bill Efcavitch was showing me data from about a year ago, which is to condense the arrays by using nanoparticles on the chips, so that instead of having a random array of poly-T tails spread across the chip, so you are looking down at a random star field — it really does look like a starry night — you are looking down at row upon row of light dots.

I'd say it's a more significant upgrade than anything they could possibly come up with. The biggest reason is that the number of data points would go up considerably, that was obvious. But the biggest challenge for the sample prep for the Helicos has been quantitation — you need to have a very good idea of how many poly-A-tailed DNA fragments you are loading into the sequencer. There are Invitrogen streptavidin chips that make it a very simple procedure if you

have enough DNA to waste a little bit on a QC like that, but with the ChIP-seq, you are pretty much flying blind for the amount. But with the condensed arrays, if they use these nanoparticles, there would be no such thing as overloading the chip, which would become a huge advantage. The sample prep would even be shorter and simpler.

### **Are you still keeping an eye out on other emerging sequencing technologies, or are you going to stick with the Helicos for now?**

Oh no, I'm totally agnostic about which platform I would ultimately be running. I'm certainly sold on single-molecule sequencers, so I am going to skip over the next-generation sequencers that are out there right now. So then there is Pacific Bio; VisiGen, which is inside Life Tech now; and I could name a couple more. I have been doing this for 30 years, so there are a few people out there who will chat with me about how all of those things are going, and what it comes down to is that I think I feel very confident writing out a check in January for a Helicos, and not thinking that some other company is going to come out with a commercially viable platform anytime soon.

I really don't think anybody is going to get their act together for three years. That doesn't really come from people telling me what other instruments are doing today, it comes from watching Helicos do what it had to do to get single-molecule sequencing to actually work, to get the optics to actually see a single molecular event, see 40 million of them occurring in just one sample, and then doing it on 50 samples at the same time. In theory, one can do that a lot of different ways, with polymerase being the thing that incorporates and spits out the color, and in theory, one can have an apparatus that can hold that object, so that you can photograph the color, but in fact, getting all of that to work, I just don't see that happening anytime soon on all those other platforms. I'll be first in line when they do come into play, but I think I'm going to be using the Helicos for quite some time before they do that.

### **What are you going to do next with the Helicos? Is it true that you are going to sequence your dog?**

Yeah, I am going to sequence my rescue dog mutt, Stella. Her mother was a collie, but we don't know who her father was. I think she definitely has got some African Basenji in her. Why would I be sequencing my dog? Because I can, and it doesn't cost me anything.

When you do a Helicos run, on each of those chips, you have to run a standard, so instead of 50 samples, you are running 48. Now that we have experience with the Helicos, we know that the only thing that the standard does, a very important job, is to help with camera focus. You absolutely positively have to have a fully functioning sample in that lane — [if your control fails], all 25 samples on that chip would fail. [So] I'm going to use my dog.

The hardest part of Stella-prep is, I've got to line up the vet to take 2 milliliters of blood from Stella because my wife won't let me do it. I've already checked, and there are no SNPs available from an African Basenji, but because I am only running two of the samples every 10 days, the project will take a while. Maybe by the time I have good deep sequencing for Stella, there will be SNPs available.

Genomeweb system

These settings are generally managed by the web site so you rarely need to consider them.

**Issue Order: 3**

